AD_____
(Leave blank)

Award Number:
W81XWH-11-1-0755

TITLE:
Adaptive Computer-Assisted Mammography Training for Improved Breast Cancer Screening

PRINCIPAL INVESTIGATOR:
Maciej Mazurowski

CONTRACTING ORGANIZATION:
Duke University,
Durham, NC 27705

REPORT DATE:
October 2013

TYPE OF REPORT:
Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

X Approved for public release; distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
October 2013	Annual	15 September 2012-14 September 2013

4. TITLE AND SUBTITLE

Adaptive Computer-Assisted Mammography Training for Improved Breast Cancer Screening

5a. CONTRACT NUMBER
W81XWH-11-1-0755

5b. GRANT NUMBER
W81XWH-11-1-0755

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Maciej Mazurowski

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University, Durham, NC 27705

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command,Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this project, we propose to research the methodology for constructing adaptive computer-aided education systems for mammography. Improved mammography education could lead to improved benefit of mammography to breast cancer care and in turn to decreased mortality from the disease. The project includes: Observer studies to collect reading data from radiology trainees; Extraction of image features (human- and computer- based); Statistical modeling of the reader data to discover patterns in their error making; Development of methodology for adaptive training that utilizes the constructed models. The proposed adaptive system could improve education in mammography. This may in turn result in improved benefit of mammography in breast cancer detection and lower mortality associated the disease.

15. SUBJECT TERMS
Mammography, radiology, education, user modeling, resident, graduate medical education

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

INTRODUCTION: In this project, we proposed to research the methodology for constructing adaptive computer-aided education systems for mammography. The project includes: Observer studies to collect reading data from radiology trainees; Extraction of image features (human- and computer- based) from mammograms; Statistical modeling of trainees reading data to discover patterns in their error making; Development of methodology for adaptive training that utilizes the constructed models. The proposed adaptive system could improve education in mammography. This may in turn result in improved benefit of mammography in breast cancer detection and lower mortality associated the disease.

BODY:

Overall progress:

Specific aim	Expected	Actual
1.1 Prepare the database of screening mammograms (year 1, months 1-6)	Completed	Completed in respect to the data for the first reader study
1.2 Obtain the approval for the human subject (observer) studies in tasks 1 and 3.	Completed	Completed
1.3 Perform an observer study in which residents will search for masses and architectural distortions (year 1, months 7-9). We expect 20 human subjects (observers) to participate in this study.	Completed	Completed
1.4 Utilize the user data collected in the observer study to develop machine learning-based individual user models (year 1, month 10 – year 2, month 6)	Completed	Main part completed

The detailed description of progress regarding each specific aim follows.

1.1 Prepare the database of screening mammograms (year 1, months 1-6)

STATUS: completed in the 2011-2012 period in respect to the data for the first reader study

1.2 Obtain the approval for the human subject (observer) studies in tasks 1 and 3.

STATUS: completed in the 2011-2012 period

1.3 Perform an observer study in which residents will search for masses and architectural distortions (year 1, months 7-9). We expect 20 human subjects (observers) to participate in this study.

STATUS: completed in the 2011-2012 period

1.4 Utilize the user data collected in the observer study to develop machine learning-based individual user models (year 1, month 10 – year 2, month 6)

STATUS: main part completed in the 2012-2013 period. Some experiments are still in progress

DETAILS:

This is the crucial aim of our grant research and this aim was our focus in the 2012-2013 period.

This task consists of developing a set of human-extracted and computer-extracted image features as well as utilizing these features to build models that predict how the trainees make errors. We initiated research in this direction.

In the 2012-2013 period, we have completed the following goals related to this aim:

- Through quantitative analysis, we have evaluated the relationship between the concepts of self-assessed case difficulty, expert assessment of difficulty, and actual resident error (this analysis was started in 2011-2012 period).
- We implemented computer vision features for analysis of mammograms, which could be used for prediction of case difficulty/error
- We conducted analysis to establish that such features can be used for prediction of false negative errors among radiology trainees. The analysis was successful.
- We collected assessments of image features from experienced radiologists
- We conducted analysis to establish whether such features can be used for prediction of false negative errors among radiology trainees. The analysis was successful.

This work has resulted in three journal manuscript submissions in 2013.

Below, we briefly present the design and results for the study that examined the relationship between difficulty and error:

Methods:

In a reader study, 7 residents as well as 3 experts interpreted 100 mammograms. For each case, the participants pointed out abnormalities (if present) and assessed difficulty of the case. We evaluated performance of the trainees in terms of area under the ROC curve (AUC), sensitivity, and specificity for low and high difficulty as assessed by a trainee himself/herself or by the experts.

Results:

Table 1 shows the distribution of positive and negative cases as assessed by residents and experts. We can see that both residents and experts assessed most cases as positive. However, notable discordance can be noted (these are errors made by trainees).

Table 2 shows the difficulty assessment by the trainees and the experts. Again, notable difference between the trainees' and the experts' assessment of difficulty is visible.

Performance of the trainees based on their own individual assessment of difficulty is shown in Table 3. One can see that statistically significantly better performance ($p=0.01$) is observed for cases that were self-assessed to be less difficult. And interesting finding is that even though the specificity drops for self-assessed more difficult cases, sensitivity actually increases.

Performance of the trainees based on expert-assessed difficulty is shown in Table 4. Again, we can see significant decrease in AUC performance as we to expert-assessed more difficult cases (p=0.001). In this case, we also significant drop is sensitivity. The difference in specificity does not reach statistical significance.

Finally, Fig. shows the performance change cased by difficulty of the case as assessed by trainees themselves and as assessed by experts. One can see that in case of expert-assessed difficulty, the change in the overall performance is predominantly explained by the change in sensitivity while for self-assessed difficulty, it is specificity that drives the change.

Conclusions:

In summary, through quantitative analysis of reader study results, we showed that both self-assessed and expert-assessed difficulty are correlated with trainee performance. We also showed that the performance change occurs mostly in terms sensitivity for expert-assessed difficulty and in terms of specificity for self-assessed difficulty.

TABLES

Table 1: Distribution of resident and attending interpretations

		Attending Interpretations		
		Positive	Negative	Total
Resident Interpretations	Positive	383	42	425
	Negative	163	112	275
	Total	546	154	700

Table 2: Distribution of resident and attending difficulty assessments

		Attending Difficulty		
		Low	High	Total
Resident Difficulty	Low	183	58	241
	High	272	187	459
	Total	455	245	700

Table 3: Resident sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) analysis by *self*-assessed difficulty.

	High (95% CI)	Low (95% CI)	Low – High (95% CI)	p-value
Sensitivity	0.707 (0.641 - 0.773)	0.614 (0.507 - 0.723)	-0.093 (-0.209 - 0.023)	0.113
Specificity	0.583 (0.456 - 0.709)	0.905 (0.850 - 0.960)	0.322 (0.203 - 0.441)	<0.001*
AUC	0.667 (0.600 - 0.735)	0.771 (0.713 - 0.829)	0.104 (0.025 - 0.182)	0.010*

CI = Confidence Intervals. * $p < 0.05$.

Table 4: Resident sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) analysis by *expert*-assessed difficulty.

	High (95% CI)	Low (95% CI)	Low – High (95% CI)	p-value
Sensitivity	0.558 (0.463 - 0.651)	0.796 (0.729 - 0.864)	0.239 (0.123 - 0.355)	<0.001*
Specificity	0.714 (0.545 - 0.884)	0.740 (0.615 - 0.865)	0.026 (-0.184 - 0.236)	0.807
AUC	0.583 (0.498 - 0.668)	0.783 (0.710 - 0.855)	0.199 (0.087 - 0.312)	0.001*

CI = Confidence Intervals. * $p < 0.05$.

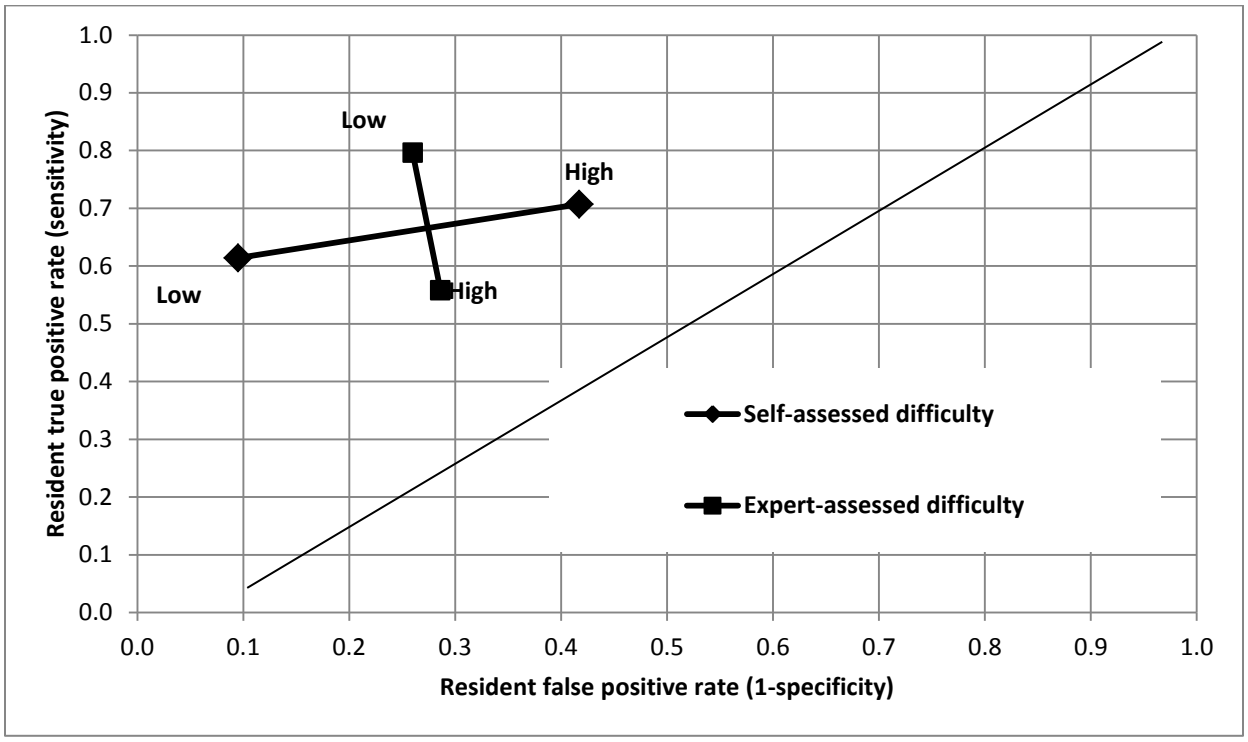


Fig. 1-Reciever operating characteristic (ROC) curve analysis for resident performance by self- and expert-assessed low and high difficulty scores.

DIFFICULTIES:

No major difficulties have been encountered as of this point. This is a crucial step of our research and the time required for this step has been somewhat longer than expected.

KEY RESEARCH ACCOMPLISHMENTS:

2011-2012

- Secured IRB approval for the study
- Retrospectively collected a set of mammograms for the study
- Conducted an observer study
- Conducted preliminary analysis of the observer study results
- Initiated development of a controlled dictionary for mammography education

2012-2013

- Evaluated the relationship between the concepts of self-assessed case difficulty, expert assessment of difficulty, and actual resident error (this analysis was started in 2011-2012 period).

- Implemented computer vision features for analysis of mammograms, which could be used for prediction of case difficulty/error and used them for prediction of false negative errors among radiology trainees
- We collected assessments of image features from experienced radiologists and established whether such features can be used for prediction of false negative errors among radiology trainees

REPORTABLE OUTCOMES:

2011-2012

- Collected a database of digital mammograms
- An extended abstract entitled "Difficulty of mammographic cases in the context of resident training: preliminary experimental data" submitted and accepted to SPIE Medical Imaging 2013 conference.

2012-2013

- Three manuscripts were submitted to journals and are currently in various review stages:
    - L. Grimm, S. V. Ghate, S.Yoon, C. M. Kuzmiak, C. Kim, **M. A. Mazurowski** (2013). 'Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features', in revision for Medical Physics, November 2013.
    - L. Grimm, C. M. Kuzmiak, S. V. Ghate, S. Yoon, **M. A. Mazurowski** (2013). 'Mammography difficulty and error making patterns in the context of resident training', submitted to Academic Radiology, October 2013
    - J. Zhang, J. Y. Lo, C. M. Kuzmiak, S. V. Ghate, S. C. Yoon, **M. A. Mazurowski** (2013), 'Using computer-extracted image features for modeling of error-making patterns in detection of mammographic masses among radiology residents', submitted to Physics in Medicine and Biology, October 2013.
- Conference proceedings paper (/extended abstract) accepted for oral presentation at SPIE Medical Imaging:
    - **M. A. Mazurowski**, J. Zhang, J. Y. Lo, C. M. Kuzmiak, S. V. Ghate, S. Yoon (2014). 'Modeling resident error-making patterns in detection of mammographic masses using computer-extracted image features: preliminary experiments', SPIE Medical Imaging 2014, in press
- Oral presentation given at SPIE Medical Imaging 2013:
    - **M. A. Mazurowski**, 'Difficulty of mammographic cases in the context of resident training: preliminary experimental data', SPIE Medical Imaging, 2013
- Conference proceedings paper published:
    - **M. A. Mazurowski**, 'Difficulty of mammographic cases in the context of resident training: preliminary experimental data', SPIE Medical Imaging, 2013

CONCLUSION:

The second year of the project has been very fruitful and resulted in three journal submissions and one conference submission. The work established the relationship between the concept of difficulty and error as well as showed that both computer-extracted and human-extracted imaging features can be used for prediction of error in detecting mammographic masses in radiology trainees.

APPENDICES: Three manuscripts submitted to international journals are appended below.

Mammography difficulty and error making patterns in the context of resident training

ABSTRACT (250 words)

Rationale and Objectives

The purpose of this study is to better understand the concept of mammography difficulty and how it affects radiology resident performance.

Materials and Methods

Seven radiology residents and three expert breast imagers reviewed 100 mammograms, consisting of bilateral medial lateral oblique and craniocaudal views, using a research workstation. The cases consisted of normal, benign, and malignant findings. Participants identified abnormalities and scored the difficulty and malignant potential for each case. Resident performance (sensitivity, specificity, and area under the receiver operating characteristic curve [AUC]) was calculated for self- and expert-assessed high and low difficulty.

Results

For cases classified by self-assessed difficulty, the resident AUC was 0.667 for high difficulty and 0.771 for low difficulty cases ($p=0.010$). Resident sensitivity was 0.707 for high and 0.614 for low difficulty cases ($p=0.113$). Resident specificity was 0.583 for high and 0.905 for low difficulty cases ($p<0.001$).

For cases classified by expert-assessed difficulty, the resident AUC was 0.583 for high and 0.783 for low difficulty cases ($p=0.001$). Resident sensitivity was 0.558 for high and 0.796 for low

difficulty cases ($p<0.001$). Resident specificity was 0.714 for high and 0.740 for low difficulty cases ($p=0.807$).

Conclusion

Increased self- and expert-assessed difficulty is associated with a decrease in resident performance in mammography. However, the etiology is due to a decrease in specificity for self-assessed difficulty but a decrease in sensitivity for expert-assessed difficulty. These trends suggest that educators should provide a mix of self- and expert-assessed difficult cases in educational materials to maximize the effect of training on resident performance and confidence.

Keywords (3-5): mammograms, education, receiver operating characteristic curves, performance

INTRODUCTION

The ability to accurately interpret mammograms is a critical skill to develop for radiology residents during training, and one clearly emphasized by the American College of Radiology and the Society of Breast Imaging education guidelines (1). Despite these stated goals, survey data from the past ten years demonstrates that even though residents are spending more dedicated time on breast imaging rotations, they are not developing sufficient confidence in their abilities and feel that only radiologists with fellowship training should routinely interpret mammograms (2-4). Reinforcing the sense of inadequacy, performance data of new radiology residency graduates demonstrates that those without subspecialty training in breast imaging require several years of on the job experience before their practice patterns fall in line with approved expectations (5). In contrast, radiologists with fellowship training in breast imaging are able to achieve desired performance goals within the first year after graduation (5). Since residents are expected to be both confident and competent in the interpretation of mammograms upon graduation from a general diagnostic residency program, there is a strong need for improved residency training in the interpretation of mammograms.

Radiology residents have reported that they find that interpreting mammograms is more stressful than interpreting other imaging studies (2), and this lack of comfort indicates that a more comprehensive understanding of the challenges facing radiology residents is needed. Prior work by multiple investigators has demonstrated several trends. Specifically, residents with less experience struggle to discriminate between benign and malignant abnormalities, resulting in a greater percentage of false-positive results (6). Residents are also less efficient in their visual search patterns as they cover more image area and struggle to differentiate true breast masses from artifacts and normal breast parenchyma (7, 8). This results in a delay in detection time and an overall increase in the time spent per mammogram (8). Despite the longer interpretation time,

residents quickly reach a time threshold beyond which they make few meaningful additional discoveries, but instead make more errors (9).

Prior work on gaze-tracking with experienced mammographers has shown that rather than utilizing inefficient "search to find" strategies favored by inexperienced readers, experts instead rely upon a holistic approach to image interpretation (10, 11). They are able to detect malignant abnormalities within seconds of seeing the image, faster than they could reasonably process all of the image information available. This ability only develops after being exposed to a large volume of mammograms as the radiologist must develop a sufficient internal reference standard. This theory is supported by cognitive processing analysis which has shown that increased expertise leads to better organized knowledge structures (Azevedo R et al. presented at the 2007 annual conference of the Cognitive Science Society). Better organized knowledge structures in turn allow for improved data-driven reasoning strategies and diagnostic planning. The end result is an increase in the number of radiologic observations and an increase in the proportion of correct diagnoses. These findings suggest that radiology residents may benefit by being shown numerous examples of normal and abnormal findings, rather than sitting through lectures which seek to emphasize the analysis of signs and features (10).

Recent work has shown that there are patterns in error making by radiology residents (12-14) (Sun S et al., presented at the 2008 SPIE Medical Imaging symposium on Image Perception, Observer Performance, and Technology Assessment). Our group has also shown (12, 13) that these patterns can be captured using statistical and machine learning models. These algorithms could then be applied to the development of personalized educational tools which specifically target an individual radiologist's weaknesses (12) (Sun S et al., 2008 SPIE Medical Imaging

symposium). Implementing such an adaptive tool during residency training may help to improve resident performance, interpretations times, and confidence.

The concept of adaptive training is highly related to the concept of error (and error-making patterns) and the concept of difficulty of mammographic cases. The purpose of this study is to better understand the concept of mammography difficulty and how it affects resident performance in order to provide additional insights that may be used in the development of educational materials. Since the concept of difficulty is subjective and dependent on the reader's expertise, we seek to discriminate the effect of self- and expert-assessed difficulty on performance.

MATERIALS AND METHODS

Reader Study

In order to quantitatively assess the relationship between difficulty assessment and error, we conducted a reader study. An Institutional Review Board protocol was obtained for the study. Seven radiology residents and 3 expert breast imaging radiologists from the Duke University Department of Radiology and the University of North Carolina at Chapel Hill Department of Radiology participated in the study. The expert interpreters were all fellowship trained in breast imaging with 7 to 14 years of experience. The residents between 4 and 12 weeks of dedicated breast imaging training. Each participant reviewed 100 mammographic cases in this reader study. Each case consisted of bilateral craniocaudal (CC) and medial lateral oblique (MLO) views. The data set consisted of normal, benign, and malignant cases. The abnormal (benign and malignant) cases contained biopsy-proven masses. Since the focus of the study was on mammographic masses, the readers were instructed to ignore calcifications.

Each participant reviewed the images on a dedicated research workstation, which closely simulated the environment of a typical diagnostic workstation. For each case, the participant clicked on any actionable abnormalities identified on each view. An abnormality was considered actionable if the reader felt that additional workup was needed, which would include additional imaging or biopsy. A timer recorded the length of time it took for the participant to complete review of the case.

After the participant finished reviewing each case, he or she was asked to report the difficulty of the case for all cases and the likelihood of malignancy for actionable cases. Each question was answered on a scale from 1 to 5. For actionable abnormalities only, malignancy was graded with a 1 considered most likely benign up to a 5 considered most likely malignant. Difficulty was graded with a 1 considered the easiest up to a 5 considered the most difficult.

Data Analysis

The expert interpretations were used to establish the ground truth. If at least two out of three experts assessed a case as actionable, then the final assessment was considered positive; otherwise, it was considered negative. We assumed that a resident made an error if his or her assessment was positive (i.e. actionable) and the ground truth was negative or when his or her assessment was negative and the ground truth was positive. While many mammography studies consider error as a difference between the reader's assessment and the biopsy results, our goal was to test how closely the resident's interpretation skills compare to expert reader's interpretations. Thus, we decided to use the expert interpretations as the ground truth for our study.

The cases were divided into low and high difficulty groups based on the residents' difficulty assessments and the experts' difficulty assessment. For each resident, to divide the cases into low and high difficulty groups based on his or her individual difficulty assessment, we found a threshold to their difficulty assessment that provided the most equal split of the 100 cases into the two groups. As a result, a case with a difficulty score of 3 for one resident may be low difficulty, but for another it may be high difficulty, depending on the distribution of lower and higher difficulty scores. We will refer to this categorization as self-assessed difficulty since the cases for each resident were divided according to his or her own difficulty assessment rather than the average assessment of other residents. To divide the cases into low and high difficulty groups based on the expert's difficulty assessment, the difficulty scores for each case were averaged across the three experts. The cases were then split into low and high difficulty groups using a threshold difficulty score that provided the most even distribution of low and high difficulty groups, which resulted in a difficulty score of 3 or greater being classified as high difficulty by the experts.

To evaluate whether self- and expert-assessed difficulty were associated with the actual performance of the residents we compared the performance of the residents for cases with high and low difficulty. We used three figures of merit: sensitivity, specificity and area under the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) was calculated using the trapezoidal rule and specifically using the *roc* function in the pROC package in R statistical software (R Development Core Team). Each of these figures of merit was calculated separately for each resident and then averaged across the residents. Furthermore we applied the Jackknife procedure (15) to find the unbiased estimate of each of the figures of merit, estimate the confidence intervals for the figures of merit as well as statistically compare the figures of merit for high and low difficulty cases. In the Jackknife procedure, one case was

removed at a time and the remaining 99 cases were used to calculate the quantity of interested. The Jackknife procedure was conducted using in-house software written in R. A p-value less than 0.05 was considered statistically significant. Initial preliminary results of this study were presented at the SPIE Medical Imaging symposium (Mazurowski MA, presented at the 2013 SPIE Medical Imaging symposium on Image Perception, Observer Performance, and Technology Assessment).

RESULTS

The distribution of resident and expert interpretations is shown in Table 1. The results demonstrate that both residents and experts classified the majority of cases as positive, 60.7% and 78.0% respectively. A breakdown of the distribution of resident and expert difficulty assessments is shown in Table 2, which shows that while the majority of resident case classifications were difficult (65.6%), the minority of expert case classifications were difficult (35.0%).

Regarding the impact of self-assessed difficulty on resident performance, with increased self-assessed difficulty there is a decrease in AUC performance ($p=0.010$), but this is due to a decrease in specificity ($p<0.001$) without a significant change in sensitivity ($p=0.113$). Self-assessed difficulty on resident performance is shown in Table 3. Regarding the impact of expert-assessed difficulty on resident performance, with increased expert-assessed difficulty there is still a decrease in performance ($p=0.001$), but this is due to a decrease in sensitivity ($p<0.001$) without a significant change in specificity ($p=0.807$). Expert-assessed difficulty on resident performance is shown in Table 4. A graphical representation of the ROC curve analysis is shown in Figure 1 which demonstrates an orthogonal relationship between self- and expert-

assessed high and low difficulty due to the differences in sensitivity and specificity between the two groups.

Residents spent an average of 51.2 seconds on self-assessed low difficulty cases and 69.3 seconds on self-assessed high difficulty cases. Residents spent an average of 60.9 seconds on expert-assessed low difficulty cases and 64.1 seconds on expert-assessed high difficulty cases.

DISCUSSION

Our data provides some important insights into the differences in difficulty perception between residents and experts, and how they affect resident performance. There is a clear disconnect between resident and attending assessment of difficulty, as evident in Table 2. Additionally, the orthogonal relationship between the effect of self- and expert-assessed difficulty seen in the ROC plane (Figure 1) demonstrates that although the difficulty assessment on overall AUC performance is similar, the impact on different aspects of performance is very different for self- and expert-assessed difficulty.

Regarding self-assessed difficulty, as residents transition from low to high difficulty cases their performance decreases (Figure 1) as they maintain their sensitivity at the cost of specificity. Residents are likely concerned about missing cancers for the cases that they themselves assessed as difficult and therefore their internal threshold for annotating a case is decreased. This results in a larger number of false positive annotations with similar sensitivity.

A different pattern of resident performance emerges for cases that experts rated as difficult. For such cases, resident performance also decreases, but this is due to a decrease in sensitivity with no significant change in specificity. This increase in false negative interpretations is likely due

to subtle masses that the experts were able to identify with difficulty, but that were missed by the residents. If the resident did not notice the mass, they might not have considered the cases difficult and as a result did not adjust their internal assessment threshold to maintain sensitivity. This likely explains the dramatic decrease in sensitivity while keeping the specificity comparable.

Our findings have important implications for developing better educational materials for residents. If faculty chose training materials solely based on their own sense of difficulty, their training materials would be biased toward cases for which the residents would exhibit low sensitivity and therefore the teaching efforts would more likely be directed toward decreasing the number of missed abnormalities. The benefits of including cases that residents themselves find difficult will be twofold. First, it will allow for the inclusion of cases for which the residents' specificity is suffering and consequently adding focus on decreasing the number of false positive calls. Second, since a major drawback of resident radiology education is a lack of confidence in mammography interpretation skills (2-4), by having residents train on cases that are self-perceived as difficult, the residents should gain more self-confidence in their abilities, as they shift cases from self-assessed high to low difficulty. These combined factors should have the greatest net effect on resident performance.

This study has some limitations. Although the study design was to closely simulate a real world experience, subject responses may have been influenced because they knew they were being studied. Individual resident mammography experience was by nature heterogeneous across the study population, which hopefully provides an accurate sample of the "average" resident, but there is no way to precisely ensure a full spectrum of resident abilities is represented. Finally, the ground truth used for this study was the expert majority opinion as we wished to assess

resident interpretation skills, but it is possible that a resident made the correct interpretation and the expert majority opinion was incorrect.

CONCLUSIONS

Our study demonstrates that the concepts of self- and expert-assessed difficulty are important to understanding resident performance during the interpretation of mammograms. It is clear that both self- and expert-assessed difficulty have different effects on the specificity and sensitivity of residents. Educators should take these differences into account when developing educational materials by selecting cases that the teacher finds difficult and ones that the student finds difficult. Consequently, this will provide a more comprehensive teaching framework, with the goal of improving the sensitivity, specificity and confidence of the residents when interpreting mammograms.

REFERENCES

1. Monticciolo DL, Rebner M, Appleton CM, et al. The ACR/Society of Breast Imaging Resident and Fellowship Training Curriculum for Breast Imaging, updated. J Am Coll Radiol 2013;10(3):207-10 e4.

2. Bassett LW, Bent C, Sayre JW, Marzan R, Verma A, Porter C. Breast imaging training and attitudes: update survey of senior radiology residents. AJR 2011;197(1):263-9.

3. Bassett LW, Monsees BS, Smith RA, et al. Survey of radiology residents: breast imaging training and attitudes. Radiology 2003;227(3):862-9.

4. Kundel HL, Nodine CF, Krupinski EA, Mello-Thoms C. Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. Acad Radiol 2008;15(7):881-6.

5. Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology 2009;253(3):632-40.

6. Nodine CF, Kundel HL, Mello-Thoms C, et al. How experience and training influence mammography expertise. Acad Radiol 1999;6(10):575-85.

7. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Acad Radiol 1996;3(12):1000-6.

8. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Acad Radiol 1996;3(2):137-44.

9. Nodine CF, Mello-Thoms C, Kundel HL, Weinstein SP. Time course of perception and decision making during mammographic interpretation. AJR 2002;179(4):917-23.

10. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology 2007;242(2):396-402.

11. Baxi SS, Snow JG, Liberman L, Elkin EB. The future of mammography: radiology residents' experiences, attitudes, and opinions. AJR 2010;194(6):1680-6.

12. Mazurowski MA, Baker JA, Barnhart HX, Tourassi GD. Individualized computer-aided education in mammography based on user modeling: concept and preliminary experiments. Med Phys 2010;37(3):1152-60.

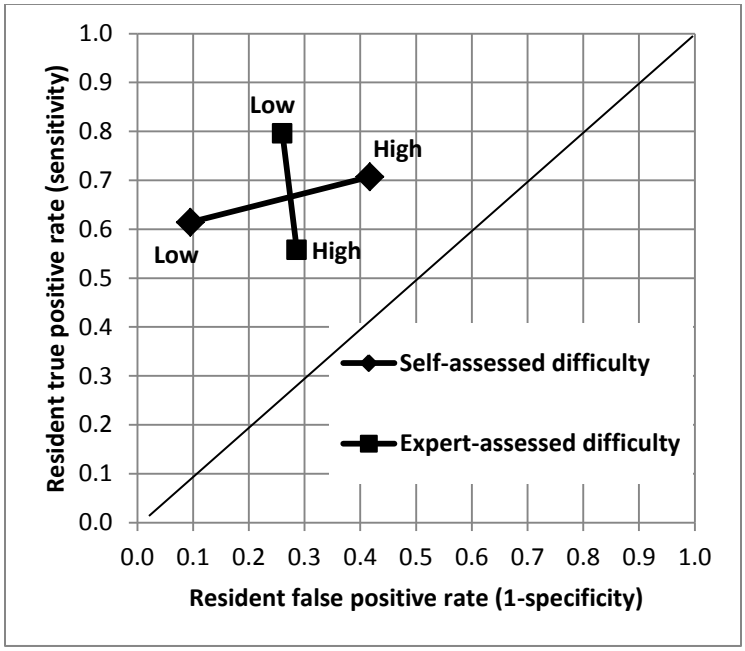
13. Mazurowski MA, Barnhart HX, Baker JA, Tourassi GD. Identifying error-making patterns in assessment of mammographic BI-RADS descriptors among radiology residents using statistical pattern recognition. Acad Radiol 2012;19(7):865-71.

14. Tourassi G, Voisin S, Paquit V, Krupinski E. Investigating the link between radiologists' gaze, diagnostic decision, and image content. J Am Med Inform Assoc 2013.

15. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*, 1st ed. Boca Raton, FL: Chapman and Hall/CRC; 1994.

FIGURES

Fig. 1-Reciever operating characteristic (ROC) curve analysis for resident performance by self- and expert-assessed low and high difficulty scores.



TABLES

Table 1: Distribution of resident and attending interpretations

		Attending Interpretations		
		Positive	Negative	Total
Resident Interpretations	Positive	383	42	425
	Negative	163	112	275
	Total	546	154	700

Table 2: Distribution of resident and attending difficulty assessments

		Attending Difficulty		
		Low	High	Total
Resident Difficulty	Low	183	58	241
	High	272	187	459
	Total	455	245	700

Table 3: Resident sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) analysis by *self*-assessed difficulty.

	High (95% CI)	Low (95% CI)	Low – High (95% CI)	p-value

Sensitivity	0.707 (0.641 - 0.773)	0.614 (0.507 - 0.723)	-0.093 (-0.209 - 0.023)	0.113
Specificity	0.583 (0.456 - 0.709)	0.905 (0.850 - 0.960)	0.322 (0.203 - 0.441)	<0.001*
AUC	0.667 (0.600 - 0.735)	0.771 (0.713 - 0.829)	0.104 (0.025 - 0.182)	0.010*

CI = Confidence Intervals. * $p < 0.05$.

Table 4: Resident sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) analysis by *expert*-assessed difficulty.

	High (95% CI)	Low (95% CI)	Low – High (95% CI)	p-value
Sensitivity	0.558 (0.463 - 0.651)	0.796 (0.729 - 0.864)	0.239 (0.123 - 0.355)	<0.001*
Specificity	0.714 (0.545 - 0.884)	0.740 (0.615 - 0.865)	0.026 (-0.184 - 0.236)	0.807
AUC	0.583 (0.498 - 0.668)	0.783 (0.710 - 0.855)	0.199 (0.087 - 0.312)	0.001*

CI = Confidence Intervals. * $p < 0.05$.

Modeling resident error-making patterns in detection of mammographic masses using computer-extracted image features

Jing Zhang[1], Joseph Y. Lo[1,3,4,5], Cherie M. Kuzmiak[2],
Sujata V. Ghate[1], Sora C. Yoon[1], Maciej A. Mazurowski[1,3]

[1] Department of Radiology, Duke University School of Medicine, Durham, NC

[2] Department of Radiology, University of North Carolina at Chapel Hill School of Medicine

[3] Duke Cancer Institute

[4] Departments of Biomedical Engineering and Electrical & Computer Engineering, Duke University, Durham, NC

[5] Medical Physics Graduate Program

Abstract:

Introduction Mammography is the most widely used screening modality for early breast cancer detection. Providing high quality mammography education to radiology trainees is essential, since good interpretation skills are needed to ensure the highest benefit of screening mammography for patients. We have previously proposed a computer-aided education system based on user models that relate human-assessed image characteristics to trainee error. In this study, as the next step in our research, we propose to build trainee models that utilize features automatically extracted from images using computer vision algorithms. A trainee model, predicts likelihood of missing each mass by that trainee. This computer vision-based approach to trainee modeling will allow for searching large databases of mammograms in order to identify difficult cases for each trainee.

Methods Our algorithm for predicting likelihood of missing a mass consists of three steps. First, a mammogram is segmented into air, pectoral muscle, fatty tissue, dense tissue, and mass using automated segmentation algorithms. Second, 43 features are extracted for each abnormality indicated by experts using computer vision algorithms. Third, error-making models (classifiers) are applied to predict the likelihood of trainees missing the abnormality based on the extracted features. The models are developed individually for each trainee using his/her previous reading data. We evaluated predictive performance of our algorithm using data from a reader study in which 10 trainees and 3 experts read 100 mammographic cases. Receiver operating characteristic methodology was applied for the evaluation.

Results The average area under the ROC curve of the error-making models for the task of predicting which masses will be detected and which will be missed was 0.607 (95% CI: 0.564-0.650). This value was statistically significantly different from 0.5 ($p<0.0001$) and therefore generally our models were able to predict which masses were detected and which were missed better than chance.

Conclusions We proposed an algorithm that was able to predict which masses will be detected and which will be missed by each individual trainee. This confirms existence of error-making patterns in detection of masses among radiology trainees. Furthermore the proposed methodology will allow for optimized selection of difficult cases for the trainees in an automatic and efficient manner.

Keywords Breast Cancer Radiology education Mammography Tumor segmentation Feature Extraction Predictive Model Random forest

1. Introduction

Numerous studies have shown that early detection with screening mammography can decrease breast cancer mortality (Siegel, Naishadham et al. 2013). However, interpretation of screening mammography can be a challenge, especially for radiology residency trainees, who may lack the experience and

judgment of sub-specialty trained radiologists. Even among experienced radiologists, studies have shown a high inter-observer variability (Elmore, Wells et al. 1994).

Prior studies have shown that with dedicated mammography education, interpretation skills and diagnostic performance can be improved (Linver, Paster et al. 1992, Leung, Margolin et al. 2007). Currently, the majority of the mammography education during residency is conducted through apprentice training. We propose that quality and efficiency of education in mammography can be improved through quantitative approach to the training. In this approach, strengths and weaknesses are identified for each resident through statistical modeling of his or her previous reading data and more optimal training protocol (e.g., consisting of more challenging cases) is developed based on these models.

Our research group has been developing an adaptive computer-aided education system in mammography. In our original study (Mazurowski, Baker et al. 2010), we proposed such system based on user modeling to capture the error making patterns of radiologists-in-training. We showed that two human-assessed features can be used to model diagnostic error in the task of distinguishing benign and malignant lesions. In our subsequent project (Mazurowski, Barnhart et al. 2012), we demonstrated that statistical pattern recognition algorithm can be successfully used to model error making in assessment of BI-RADS. Finally, we explored the use of collaborative filtering algorithms for modeling of resident errors in mammography (Mazurowski and Tourassi 2011).

Besides our group's research, some initial studies on constructing an educational system that analyzes the differences between radiologists were presented by Sun *et al*. (Sun, Taylor et al. 2008, Sun, Taylor et al. 2008). Their focus was largely on developing related ontology for training systems. Furthermore, while not focused on radiology education, some other general studies on human perception and error making have relevance to our work. For example, in (Mello-Thoms, Dunn et al. 2002) , Mello-Thoms *et al.* related areas on mammograms that attracted visual attention with their spatial frequency representations. Tourassi *et al.* (Tourassi, Voisin et al. 2013) investigated the link among image content, human perception, human cognition, and human error for mammographic breast cancer detection. Additionally, Voisin *et al.* (Voisin, Pinto et al. 2013) has investigated relationship of diagnostic error with eye gaze and imaging features (human- and computer-extracted). However, the focus of the study is on eye gaze tracking and no evidence of independent predictive accuracy in predicting error is presented for automatically extracted imaging features. Finally, while prior computer-aided diagnosis literature (Cheng, Shi et al. 2006), (Tang, Rangayyan et al. 2009), (Oliver, Freixenet et al. 2010) has a significantly different focus, some of the image processing features proposed in that field are of use for our purpose.

In this paper, we present the next step in the development of an adaptive mammography education system. For each individual trainee, we identified previous error patterns and based on these patterns, built a model of error-making behavior. Such a model is potentially useful for predicting future errors in detecting or missing lesions. A major innovation of our study is that, instead of using the mass features provided by radiology experts, the proposed method uses a variety of automatically extracted image features to analyze properties of masses which might be predictive of difficulty. We propose that by predicting error making patterns during training, residency programs may potentially improve educational outcomes of trainees. This may be accomplished by focusing on teaching a selection of cases deemed most "difficult" for individual residents.

2. Methods

The proposed algorithm for prediction of likelihood of making errors proceeds as follows: First a mammography image is segmented into different parts using image processing algorithms previously developed for computer-aided detection (CAD) systems. Then, features are extracted to describe the properties of each lesion and its context. Finally, a classifier is applied to predict the likelihood of the lesion being missed using extracted features. Fig. 1 illustrates the flowchart of the proposed algorithm.

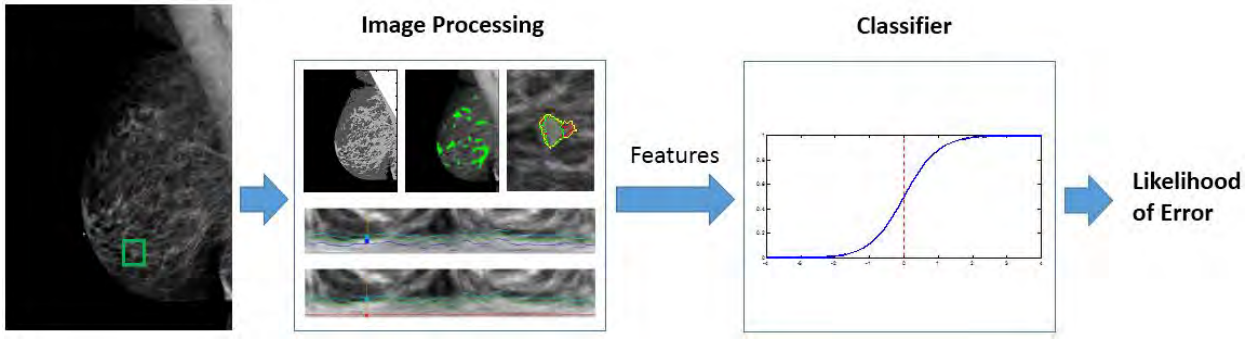


Fig. 1 Flowchart of the proposed algorithm

This section describes our methodology in detail including the reader study, the definition of error used in our study, the algorithms used for image segmentation, feature extraction, and modeling the relationship between the extracted features and error. We also describe how we evaluated our algorithm using the data from the reader study to establish whether the predictions of likelihood of making error provided by our algorithm are accurate.

2.1 Reader study

We conducted a reader study in which 3 expert radiologists (breast imaging attendings) and 10 trainees interpreted the same set of 100 full field digital mammography cases collected at Duke University Medical Center. Among the 10 trainees, 7 were radiology residents and 3 were novices (2 medical imaging researchers and 1 medical student) representing pre-residency trainees. Institutional Review Board approval was secured for this study. The set of cases used contained a mixture of negative, benign, and malignant (biopsy proven) cases. For each case, 4 views were available to the readers: LCC, LMLO, RCC, and RMLO. Prior mammograms were not available. The readers were instructed to locate the actionable abnormalities (if present) and assign the likelihood of malignancy. Since the focus of the study was on mammographic masses, the readers were instructed to ignore microcalcifications.

2.2 Definition of error

Locations of the masses was determined by 3 expert readers. If there were two or more marks given to a certain location, we considered this location a mass. A threshold T_d, was used to determine whether two marks from two different experts belonged to the same location. The centroid of all points for one location was considered as the centroid of the mass. The T_d was calculated by dividing the average radius of the breast masses (9mm) (Timp, Karssemeijer et al. 2003) by the pixel spacing of the images in the data set (0.0941 mm): T_d = 9mm/0.0941mm=96 pixels. The same distance criterion (9 mm from the centroid of the mass) was applied for trainees to determine whether a mass has been detected.

The first row of Fig. 2 illustrates an example of a mammogram with one mass found by the experts using the criteria described above. Locations identified by the trainees for the same mammographic image are shown in the second row of Fig. 2. Six out of ten trainees detected this abnormality and there is a non-mass region marked as mass by a trainee.

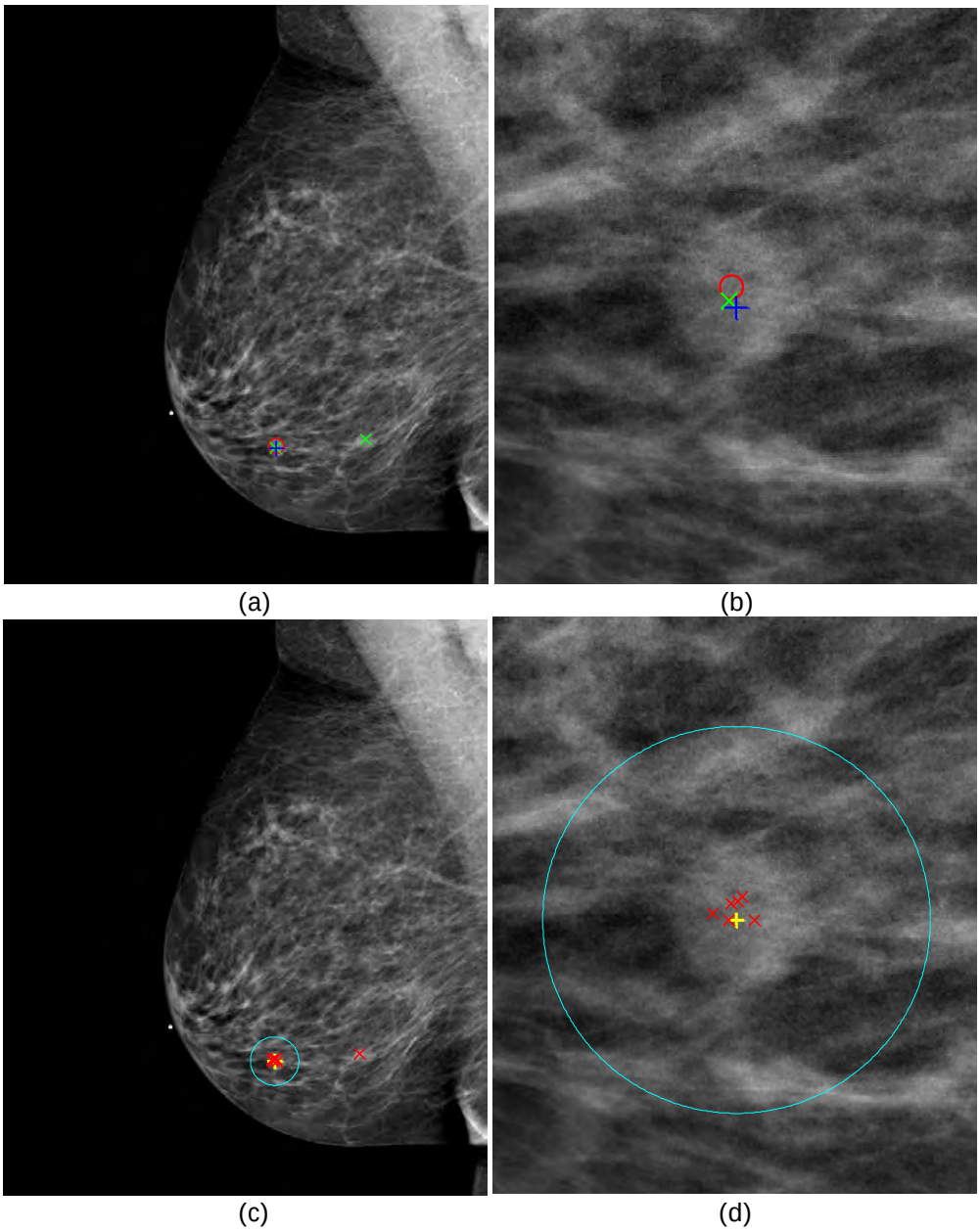


Fig.2 Reader study and error definition. (a) The blue cross, the red circle, and the green X-mark are the points assigned by three experts. (b) The close-up of the mass region in (a). (c) Locations indicated by trainees. The red X-marks are the locations indicated by 10 trainees; the yellow cross is the centroid of the mass as indicated by the experts, and the cyan circle shows the round region centered at yellow cross with the radius 96 pixels. Six out of ten trainees detected this abnormality (i.e. there are six red x-marks inside the cyan circle) and a non-mass region was marked as mass by a trainee (i.e. there is a red x-mark outside the cyan circle). (d) The close-up of the mass region in (c).

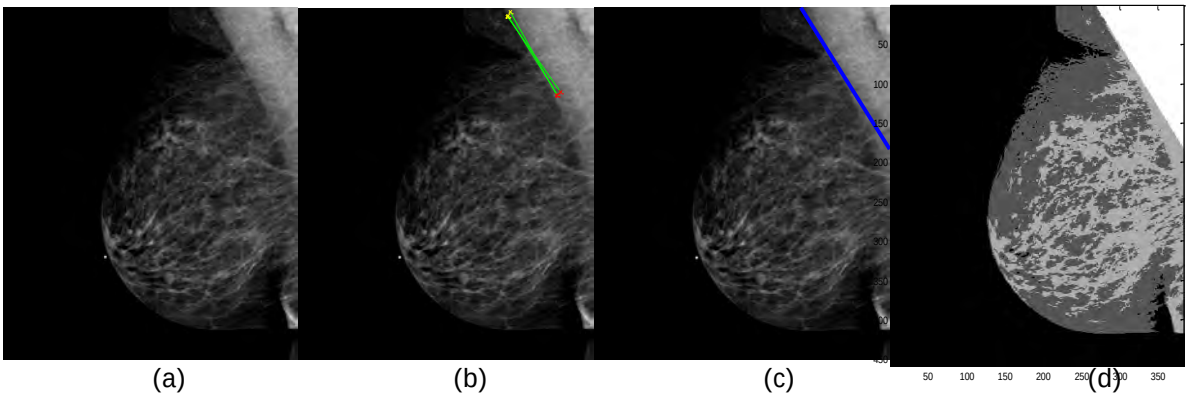
2.3. Image Segmentation and Feature Extraction

In this subsection, we describe the image segmentation and feature extraction in our approach. The image features extracted from the segmented mammography images were used to describe the characteristics of masses and breasts, and then to predict whether the trainees detected the mass or not.

2.3.1 Segmentation

The segmentation process included: (1) Mammography image segmentation into air, pectoral muscle, fatty tissue, and dense tissue; (2) Mass region segmentation to find the boundary of the mass based on the marks provided by the experts; (3) Suspicious region segmentation to detect suspicious regions that resemble masses in the image.

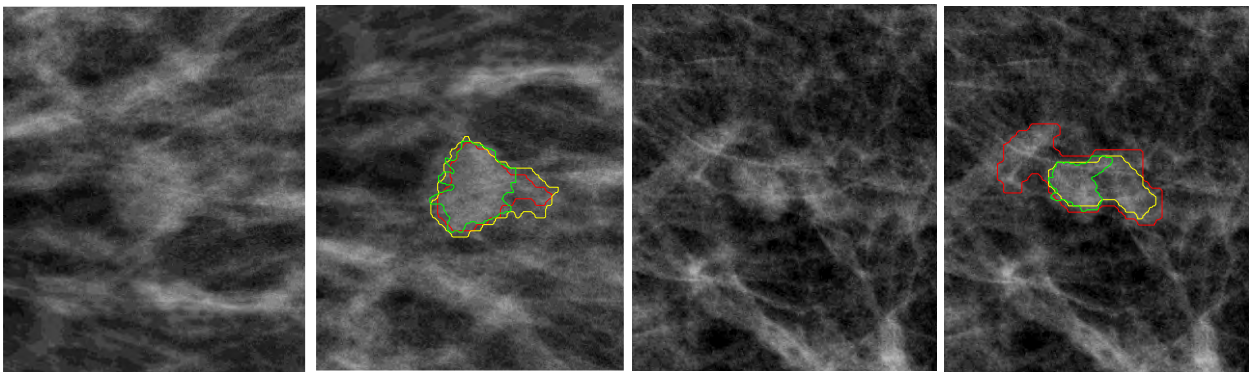
The first step to segment mammography image into air, pectoral muscle, fatty tissue, and dense tissue, consisted of us locating the valley in the mammographic image intensity histogram to identify skin-air interface. Then, after applying two directional gradient filters (+45 degree and -45 degree) on LMLO and RMLO images separately, we used Hough Transform to find straight lines and computed best-fit line to approximate pectoral muscle edge. Finally, we used Gaussian Mixture Model (GMM) method to separate fatty and dense tissues based on their intensities. Fig. 3 illustrates this segmentation step.



(a) (b) (c) (d)

Fig. 3 Pectoral muscle edge detection. (a) Original image. (b) The green straight lines with end points in yellow and red are computed by Hough Transform. (c) Pectoral muscle edge computed by finding the best fit line. (d) Breast region was segmented into air (black), fatty tissue (dark gray) and dense tissue (light gray), and pectoral muscle (white).

In the second step, the mass boundary was identified using two methods. The first method was dynamic programming-based rubber band method (Timp and Karssemeijer 2004). This method transforms a round image region centered at the mass centroid with radius R to a polar coordinate system. Then, the gradient, size, and intensity information extracted from the image in the polar system are combined to form a cost matrix. Finally, dynamic programming is used to find a path with the lowest cost in the cost matrix and this path is used as the boundary of the mass. The second mass segmentation method was region growing (Adams and Bischof 1994). Given a seed point determined by the experts, this method computes the similarities between the seed region and its neighboring pixels. If the similarity is smaller than a predetermined threshold, the mass region is grown by including its neighboring pixels. The method works iteratively and stops when no new pixels can be included. We adopted two seed strategies for region growing method: One with a fixed seed region and the other with an adaptive seed region which is updated at each iteration. By comparing the three segmentation results, we are able to learn the segmentation difficulty of a case. Generally, the segmentation methods provide similar boundaries for easy cases and rubber band method can provide more accurate boundaries than region growing methods for hard cases. Fig 4 illustrates two masses with different segmentation difficulties. We can see that the methods provided similar boundaries for the first mass with lower segmentation difficulty, and quite different boundaries for the second mass with higher segmentation difficulty.



(a) (b) (c) (d)

Fig. 4 Identification of mass boundary. (a) and (c) show original masses. (b) and (d) show the segmentation results. The green, yellow, and red boundaries are the mass regions detected by rubber band method, region growing method with fixed seed region, and region growing method with adaptive seed region.

In the third step, a suspicious region was defined as mass-resembling regions. We detected suspicious regions in the image using multi-scale Frangi filtering method (Frangi, Niessen et al. 1998), which uses all eigenvalues of Hessian matrix (Thacker 1989) to examine local structure. Fig. 5 shows the suspicious regions detected in an example mammogram in green.

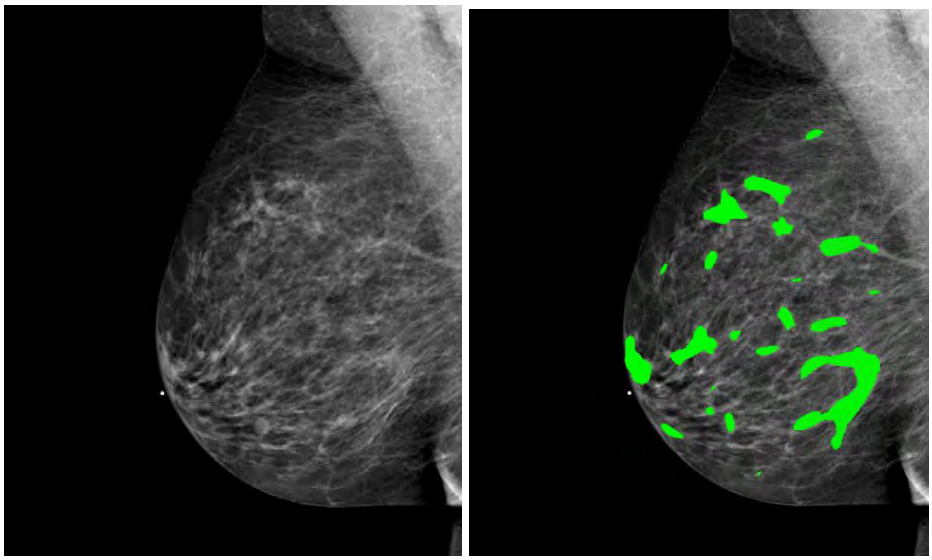


Fig. 5 Suspicious region segmentation. The green regions are suspicious regions detected by multi-scale Frangi filtering method.

2.3.2 Feature Extraction

We extracted 43 features for each mass using the regions segmented in subsection 2.3.1. We divided the features into three categories: mass-based features, context-based features, and symmetry-based features. The features captured qualities of masses and the image that we suspected might be related to likelihood of making error by trainees. Some of the features presented below are similar to each other. We decided to extract a large number of potentially correlated features and identify ones that are the most important in the modeling step.

Table 1 lists all features extracted for the three categories. The features are denoted as F1 through F43. Please note that we use 'region growing I' and 'region growing II' to represent the region growing

method with the fixed seed region and the region growing method with adaptive seed region respectively. Below, we describe the features that require more detailed explanation:

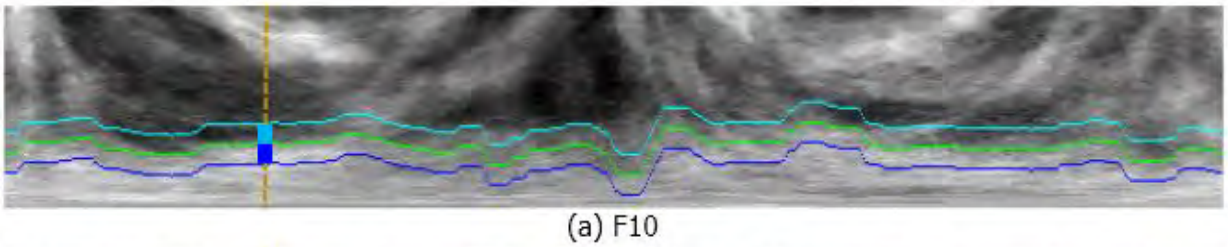
(1) Mass-based features: The 17 features in this category are extracted to describe the characteristics of a mass itself. F4 indicates the normalized mass intensity computed as below:

$$Normalized\ Mass\ Intensity = \frac{Mass\ Intensity - Fat\ Tissue\ Intensity}{Dense\ Tissue\ Intensity - Fat\ Tissue\ Intensity}$$

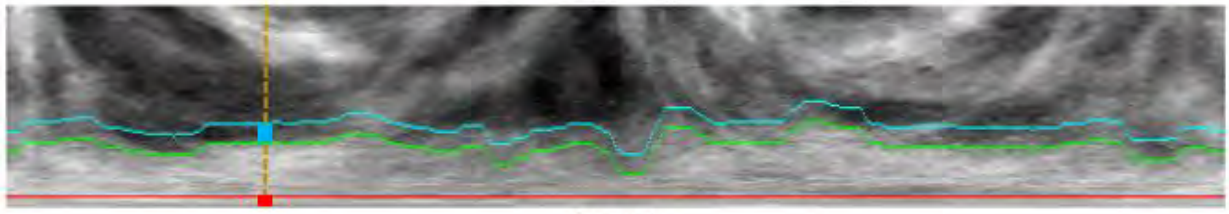
F6 shows the spiculation of a mass by measuring mass solidity using the following equation:

$$Mass\ Solidity = \frac{Mass\ Area}{Area\ of\ Minimum\ Convex\ Hull\ containing\ mass}$$

Both F8 and F9 measure the strength of the mass boundary. F8 is the total cost of the boundary pixels provided by the rubber band method and F9 is the mean gradient of mass boundary pixels. Because only two neighboring pixels are used for gradient computation, F9 is not robust to the noise. We extended it by computing F10 and F11. F10 is defined as the number of angles, along the angle axis of the polar coordinate system, whose mean intensity differences between N consecutive pixels outside and inside the boundary are bigger than (*mass intensity – fatty tissue intensity*)/2 over the total number of angles (i.e. 360). N=10 was used in our experiments.). F11 is computed similarly to F10 but the N consecutive pixels inside the boundary in F10 are replaced by the pixels that are less than M pixels to the mass centroid. M=5 was used in our experiments. Fig. 6-a and 6-b illustrate the mass region shown in Fig. 4-a in the polar coordinate system. F10 computes the intensity difference of thick vertical cyan and blue lines and F11 computes the intensity difference of thick vertical cyan and red lines. Furthermore, two similar features, F12 and F13, are proposed to analyze the intensity changes inside the mass region. F12 is the number of angles that have at least one pixel inside the mass region whose the intensity difference between N consecutive pixels above and below this pixel is bigger than (*mass intensity – fatty tissue intensity*)/2 over the total number of angles (i.e. 360). F13 is computed similarly to F12 but the N consecutive pixels below the pixel in F12 are replaced by the pixels that are less than M pixels to the mass centroid. Therefore, F10 to F13 can describe whether the detected mass region connects with fatty tissue. F16 and F17 indicate the location of a mass. F16 is the distance value at the mass centroid of the normalized distance transformed breast region. It indicates the shortest distance from the mass centroid to breast edge. F17 is the y coordinate of mass centroid divided by the y coordinate of the point on the breast edge that is furthest to the chest edge of the mammogram. Fig. 7 shows the distance transform of the breast in Fig. 3-a. F16 is d_c and F17 is y_c/Y_m.



(a) F10



(b) F11

Fig. 6 Feature F10 and F11. The figures show the representation of the mass region in polar coordinates. (a) The green curve is the mass boundary. The cyan and blue curves are the boundaries of N consecutive pixels outside and inside the mass boundary. F10 computes the intensity difference between the thick vertical cyan and blue lines (b) The green curve is the mass boundary. The cyan is the boundaries of N consecutive pixels outside the mass boundary. The red horizontal line is the boundaries

of pixels that are less than M pixels to the mass centroid. F11 computes the intensity difference between the thick vertical cyan and red lines.

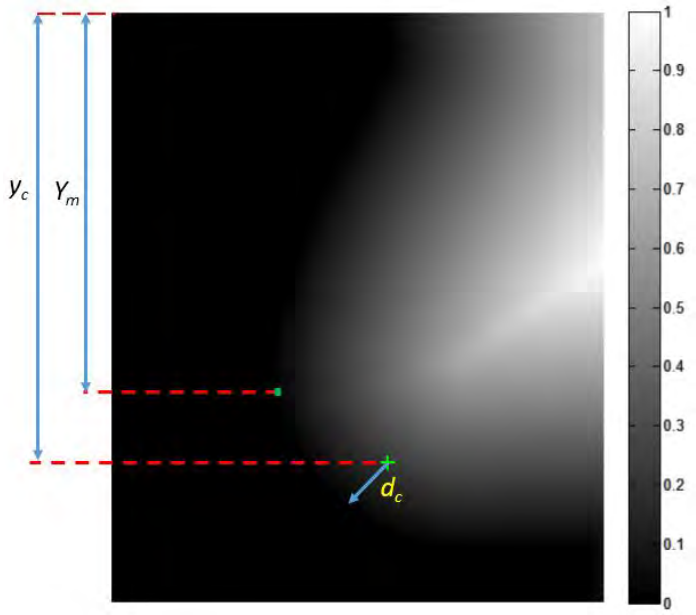


Fig. 7 Feature F16 and F17. The figure shows the normalized distance transform of the breast region. The value of each pixel indicates the normalized shortest distance from that pixel to breast edge. F16 is d_c that indicates the shortest distance from the mass centroid (green plus) to breast edge. F17 is y_c/Y_m, which is the y coordinate of mass centroid divided by the y coordinate of the point on the breast edge that is the furthest from the chest edge of the mammogram (green dot).

The rationale behind extracting features in this category is that the appearance of a mass is highly related to the likelihood of finding it. For example, we expect that the masses with high intensity, round shape, and sharp edge are easily identified by the trainees.

(2) Context-based features: The 23 features in this category describe the relationships between the mass and other regions in breast. F19 to F20 indicate the complexity and the number of holes of dense tissue. F24 to F32 are all computed using the suspicious regions, excluding the one that overlaps with the mass, detected in Section 2.3.1. F27 to F29 measure the similarities among the mass and the suspicious regions. They are computed as follows: the normalized sizes and intensities of the mass and eligible suspicious regions (radius 5-15mm) are used to compute Euclidean distance between each suspicious region and the mass to indicate their similarities. Based on the obtained similarities, F27, F28, and F29 record the average of all similarity values, the biggest similarity value, and the number of suspicious regions whose similarities are bigger than a threshold T_s, respectively. F30 to F32 are calculated similarly to F27 to F29, but besides intensity and size, suspicious region solidity is also used for similarity computation. A round area C centered at the mass centroid with radius R (the radius used in the rubber band method) is used to investigate the relationship between the mass and its neighboring area. F33 and F34 indicate the number of suspicious regions inside C and the area ratio between the mass and the suspicious regions inside C. F35 is local dense tissue density inside C.

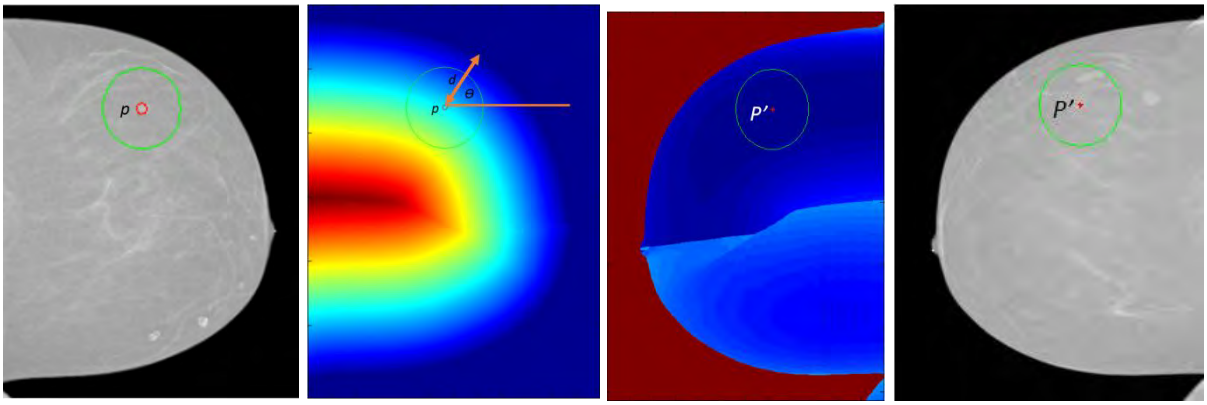
$$Local\ Dense\ Tissue\ Density = \frac{Area\ of\ Dense\ Tissue\ within\ Neighboring\ Region}{\pi * R * R}$$

F36 is the Frangi filtering response at the mass centroid. F37 computes the intensity ratio of the following two regions: the mass region and its surrounding region with the same size obtained by morphological dilation operation. F38 computes the area under curve (AUC) of a receiver operating characteristic (ROC) curve assuming that intensity of pixels are the predictors and pixels within the mass and outside of the mass belong to two different classes.

The rationale for extracting features in this category is that the intensity and distribution of dense tissue and the suspicious regions with high similarity to masses can affect the likelihood of finding the masses. For example, the mass that overlaps with dense tissue or is neighbored by many similar-looking regions is much more likely to be missed than the mass located in a clear region. We use the features in this category to capture this context information of a mass and indicate its difficulty level.

(3) Symmetry-based features: The 3 features in this category describe the symmetry of two breasts around the mass region using two image pairs (LCC and RCC, LMLO and RMLO).

Before extracting symmetry features, we need to find the corresponding regions in an image pair. We first use two location features F17 and F18, and the gradient direction Θ at the mass centroid P in distance transformed image to generate a feature space. Therefore, the mass centroid can be expressed as a point in the feature space (d, Y, Θ), where d is F17 and Y is F18. After that, we compute the same feature space for the breast region in the other image. The point with the shortest Euclidean distance to P is marked and is the corresponding point P' in the other breast image. Fig 8 shows the process to find corresponding regions in a CC image pair.



(a) (b) (c) (d)

Fig. 8 The process to find corresponding regions in a CC image pair. (a) is a LCC image with the mass centroid (red circle) P and its neighboring area within the green circle. (b) shows the distance transformed image and gradient direction Θ at the mass centroid. (c) is the feature space of the RCC image and the point with smallest Euclidean distance to P is marked as the corresponding point P'. (d) shows the corresponding region in the RCC image.

F41 is the mean intensity difference between two corresponding regions. F42 indicates the intensity difference after normalization using the intensities of fatty tissue and dense tissue. F43 is the dense tissue difference by subtracting F35, local dense tissue densities, of two images.

The rationale for extracting features in this category is to simulate the mass identification process of trainees, who usually compare LCC with RCC and LMLO with RMLO and find masses based on the asymmetry of two images. The features in this category can capture symmetry information by finding the corresponding area and indicate the difficulty level of mass.

Table 1. Extracted Features

Feature category	Feature description
Mass-based	F1: Mass Area
	F2: Intensity of Mass Centroid
	F3: Mean Intensity of Mass Region
	F4: Normalized Intensity of Mass Region
	F5: Mass Circularity
	F6: Mass Solidity
	F7: Standard Deviation of the Intensities of Mass

	F8: Mass Rubber Band Cost
	F9: Mass Boundary Gradient
	F10: Directional Intensity changes along Mass Boundary Band
	F11: Directional Intensity changes between Mass Boundary and Mass Centroid
	F12: Directional Intensity changes within Mass Region
	F13: Directional Intensity changes between Mass Region and Mass Centroid
	F14: Mass Area (Rubber Band) / Mass Area (Region Growing I)
	F15: Mass Area (Rubber Band) / Mass Area (Region Growing II)
	F16: Mass Location X
	F17: Mass Location Y
Context-based	F18: Breast Density
	F19: Dense Tissue Solidity
	F20: Dense Tissue Euler Number
	F21: Mass Area (rubber band) / Dense Tissue Area
	F22: Mass Area (region growing I) / Dense Tissue Area
	F23: Mass Area (region growing II) / Dense Tissue Area
	F24: Number of Suspicious Regions
	F25: Mean Intensity of Suspicious Regions
	F26: Suspicious Region Area
	F27: Number of Suspicious Regions based on Similarities of Intensity and Size
	F28: Maximum Similarity to suspicious regions based on Intensity and Size
	F29: Average Similarity to suspicious regions based on Intensity and Size
	F30: Number of Suspicious Regions based on Similarities of Intensity, Size, and Shape
	F31: Maximum Similarity to suspicious regions based on Intensity, Size, and Shape
	F32: Average Similarity to suspicious regions based on Intensity, Size, and Shape
	F33: Number of Local Suspicious Regions
	F34: Local Suspicious Region Area Ratio
	F35: Local Breast Density
	F36: Frangi Filtering Response at Mass Centroid
	F37: Mean Intensity of Mass Region / Mean Intensity of Neighboring Area
	F38: Intensity-based area under the ROC curve
	F39: Is Mass in Suspicious Region
	F40: Is Mass in Pectoral Muscle
Symmetry-based	F41: Intensity Difference
	F42: Normalized Intensity Difference
	F43: Dense Tissue Difference

2.4 Predictive modeling

Given a set of the features extracted in section 2.3.2 for all masses, we developed predictive models that used the image features as input (independent variables) and the occurrence of error as output (dependent variable). The actual statistical modeling was preceded by two simple filtering/feature selection steps. Specifically, we removed features such that 95% or more of the values of the feature were identical and then removed features that had very high correlation with other features ($r>=0.95$). Following the "data cleaning" steps, we constructed the classification models using multivariate logistic regression. We used the *mnrfit* function in MATLAB (Mathworks, Natick, MA). Only training data was used for the feature selection and classifier development (please see details on cross validation in section 2.5).

After the model was developed using training data, likelihood of missing the mass was predicted for each of the masses in the testing data. This prediction was enhanced with the following adjustment: If there was only one abnormality marked by the experts in each view of one breast, we assumed that the marks correspond to the same abnormality. Therefore if two different likelihoods of being missed by a resident in question were predicted by the classifier, we used the lower value for both marks. Our rationale was that if a mass appears more clearly in one view and a trainee will find it in that view, the likelihood of missing the same abnormality in the other view drastically decreases as compared with prediction of difficulty made by the other view alone since the trainee is aware of its existence.

2.5 Evaluation

In our evaluation, we simulated a situation in which a set of prior interpretation data (i.e. indicated location for prior mammograms) is available for a trainee and a prediction needs to be made regarding likelihoods of missing specific lesions for cases that have not been previously seen by the trainee. The details of our evaluation follow.

To calculate predictions of likelihood of error for each mass, we used the leave-one-case-out cross-validation [19] approach. Specifically, following this approach, one case (with potentially multiple masses) is excluded from the dataset and the remaining cases are used for development of the predictive model. Then the developed predictive model is used to predict likelihood of error for the masses in the one excluded case. The model development and prediction process is repeated multiple times such that each case is excluded from model development and assigned likelihood of error (separately for each abnormality within the case) once. The leave one-case-out process was repeated for each trainee individually.

To evaluate the performance of the models, we calculated the area under the ROC curve individually for each trainee for evaluation purpose. Please note that the task of our classifier is to determine whether a particular mass was detected by the trainee of interest rather than whether it corresponds to a malignant abnormality. Therefore a true positive classifier decision corresponds to the situation when the classifier assessed the mass as positive (detected) and the mass was detected by the trainee. True negative corresponded to the situation when the classifier considered the mass negative (not detected) and the resident did not find the mass. True negative classifier decisions mean that the classifier was able to correctly predict the trainee's error.

Furthermore, we applied the jackknife procedure (Efron and Tibshirani 1993) to estimate the confidence interval for the average AUC as well as test whether the mean AUC is significantly higher than 0.5. In the Jackknife procedure, one case was removed at a time and the remaining cases were used to calculate the AUCs for each of the trainees and then average them to obtain the final figure of merit. This is repeated n times (where n is the number of observations) and the obtained n AUC estimates are used to calculate confidence interval as well as p-value for difference between average AUC and 0.5. We applied in-house software written in R to execute the Jackknife procedure.

3 Experimental results

Based on the criteria described in the reader study section (section 2.1), the experts localized 153 abnormalities in the 400 mammography images. The number of detected masses, along with corresponding sensitivity for each trainee is presented in Table 2. A notable variability in sensitivity can be observed.

Table 2. Sensitivities and false positives of trainees

Trainee	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Number of detected Masses	91	67	103	98	61	106	93	63	74	72
Sensitivity	0.595	0.438	0.673	0.641	0.399	0.693	0.608	0.412	0.484	0.471

Regarding the models of error-making, the average cross validation performance of the models in terms of AUC was 0.607 (95% CI: 0.564-0.650). The average performance was statistically significantly higher than 0.5 ($p<0.0001$) which means that that in general the models were able to distinguish (better than chance) masses that will be detected from those that will be missed. The performance of the individual models for each trainee varied from 0.473 to 0.684 showing variability in effectiveness of the error-making models for individual trainees. The AUC performance of the models for each trainee is shown in Fig. 9.

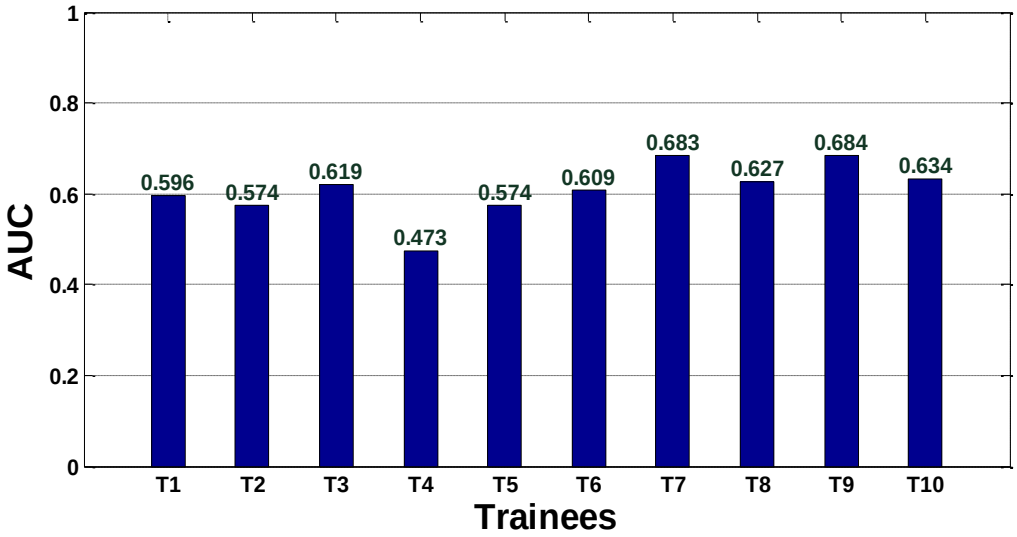


Fig. 9 Cross validation performance of the error-making models for each of the trainees

4 Conclusions and Discussion

In this study, we have taken a significant step towards the development of an adaptive computer-aided education system. Previously, we have shown that human-extracted imaging features can be used to predict error made by trainees. In this study, we applied computer vision algorithms to extract features which can be used for error prediction. Our results not only demonstrate the existence of the error-making patterns among residents when performing the task of detecting mammographic masses, but also that the developed methodology can be directly applied to a dataset of mammographic cases in order to select cases that will pose higher difficulty to the trainee in question.

While the seemingly low average area under the ROC curve of 0.607 obtained by the error-making model would not be acceptable for some other applications (such as diagnosis of mammographic masses where AUCs reach 0.9), it may be highly useful in the context of radiology education. In this context, any guidance on which cases are more likely to pose a problem to the trainees could increase the efficiency of the training by enriching the educational material with cases challenging for the resident. A predictive model with AUC values that are statistically significantly higher than 0.5 (such as our model) provides such guidance. Furthermore, such value is quite expected since rather than predicting malignancy of an abnormality, the proposed models predict behavior of human readers which can be affected by factors difficult to measure such as fatigue, surrounding distractions, or anxiety related to the experiment. Future studies will aim at improving the prediction accuracy of the error-making models as well evaluation of the benefit of presentation of educational material enriched by cases predicted to be difficult.

While our experimental results confirm existence of error-making patterns in the context of detecting of mammographic masses, the most significant advancement provided by our study lies in application of

automatically extracted features for prediction of likelihood of error. In our approach, we combine a thorough analysis of the abnormality and its context using computer vision with statistical modeling of error-making patterns. Since our approach does not require expert radiologist to assign features to the abnormalities and the cases, but only requires pointing out the abnormality (single click), it will allow training programs for searching larger databases of mammograms in order to identify cases that are the most educationally useful to the trainees. Such an approach may potentially allow training programs to identify and address cases which are challenging to each individual resident, thereby improving educational performance.

References:

Adams, R. and L. Bischof (1994). "Seeded region growing." Pattern Analysis and Machine Intelligence, IEEE Transactions on **16**(6): 641-647.

Cheng, H., X. Shi, R. Min, L. Hu, X. Cai and H. Du (2006). "Approaches for automated detection and classification of masses in mammograms." Pattern recognition **39**(4): 646-668.

Efron, B. and R. Tibshirani (1993). An introduction to the bootstrap, CRC press.

Elmore, J. G., C. K. Wells, C. H. Lee, D. H. Howard and A. R. Feinstein (1994). "Variability in radiologists' interpretations of mammograms." New England Journal of Medicine **331**(22): 1493-1499.

Frangi, A. F., W. J. Niessen, K. L. Vincken and M. A. Viergever (1998). Multiscale vessel enhancement filtering. Medical Image Computing and Computer-Assisted Interventation—MICCAI'98, Springer**:** 130-137.

Leung, J. W., F. R. Margolin, K. E. Dee, R. P. Jacobs, S. R. Denny and J. D. Schrumpf (2007). "Performance parameters for screening and diagnostic mammography in a community practice: Are there differences between specialists and general radiologists?" American Journal of Roentgenology **188**(1): 236-241.

Linver, M., S. Paster, R. Rosenberg, C. Key, C. Stidley and W. King (1992). "Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases." Radiology **184**(1): 39-43.

Mazurowski, M. A., J. A. Baker, H. X. Barnhart and G. D. Tourassi (2010). "Individualized computer-aided education in mammography based on user modeling: Concept and preliminary experiments." Medical physics **37**: 1152.

Mazurowski, M. A., H. X. Barnhart, J. A. Baker and G. D. Tourassi (2012). "Identifying error-making patterns in assessment of mammographic BI-RADS descriptors among radiology residents using statistical pattern recognition." Academic Radiology **19**(7): 865-871.

Mazurowski, M. A. and G. D. Tourassi (2011). Exploring the potential of collaborative filtering for user-adaptive mammography education. Biomedical Sciences and Engineering Conference (BSEC), 2011, IEEE.

Mello-Thoms, C., S. Dunn, C. F. Nodine, H. L. Kundel and S. P. Weinstein (2002). "The perception of breast cancer: what differentiates missed from reported cancers in mammography?" Academic radiology **9**(9): 1004-1012.

Oliver, A., J. Freixenet, J. Marti, E. Pérez, J. Pont, E. R. Denton and R. Zwiggelaar (2010). "A review of automatic mass detection and segmentation in mammographic images." Medical Image Analysis **14**(2): 87-110.

Siegel, R., D. Naishadham and A. Jemal (2013). "Cancer statistics, 2013." CA: a cancer journal for clinicians **63**(1): 11-30.

Sun, S., P. Taylor, L. Wilkinson and L. Khoo (2008). Individualised training to address variability of radiologists' performance. Medical Imaging, International Society for Optics and Photonics.

Sun, S., P. Taylor, L. Wilkinson and L. Khoo (2008). An ontology to support adaptive training for breast radiologists. Digital Mammography, Springer**:** 257-264.
Tang, J., R. M. Rangayyan, J. Xu, I. El Naqa and Y. Yang (2009). "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances." Information Technology in Biomedicine, IEEE Transactions on **13**(2): 236-251.
Thacker, W. C. (1989). "The role of the Hessian matrix in fitting models to measurements." Journal of Geophysical Research: Oceans (1978–2012) **94**(C5): 6177-6196.
Timp, S. and N. Karssemeijer (2004). "A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography." Medical Physics **31**: 958.
Timp, S., N. Karssemeijer and J. Hendriks (2003). Analysis of changes in masses using contrast and size measures. Digital Mammography, Springer**:** 240-242.
Tourassi, G., S. Voisin, V. Paquit and E. Krupinski (2013). "Investigating the link between radiologists' gaze, diagnostic decision, and image content." Journal of the American Medical Informatics Association.
Voisin, S., F. Pinto, G. Morin-Ducote, K. B. Hudson and G. D. Tourassi (2013). "Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography." Medical Physics **40**: 101906.

Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features

ABSTRACT (500 words)

Purpose:

The purpose of this study is to explore BI-RADS features as predictors of individual errors made by trainees when detecting masses in mammograms.

Methods:

Ten radiology trainees and three expert breast imagers reviewed 100 mammograms consistent of bilateral medial lateral oblique and craniocaudal views on a research workstation. The cases consisted of normal and biopsy proven benign and malignant masses. For cases with actionable abnormalities, the experts recorded breast (density and axillary lymph nodes) and mass (shape, margin, and density) features according to the BI-RADS lexicon, as well as the abnormality location (depth and clock face). For each trainee, a user-specific multivariate model was constructed to predict the trainee's likelihood of error based on BI-RADS features. The performance of the models was assessed using area under the receive operating characteristic curves (AUC).

Results:

Despite the variability in errors between different trainees, the individual models were able to predict the likelihood of error for the trainees with a mean AUC of 0.611 (range: 0.502 – 0.739, 95% Confidence Interval: 0.543 – 0.680, $p < 0.002$).

Conclusions:

Patterns in detection errors for mammographic masses made by radiology trainees can be modeled using BI-RADS features. These findings have potentially important implications for the development of future educational materials that are personalized to individual trainees.

KEYWORDS (5 words): mammography, education, receiver operating characteristic curves, performance

INTRODUCTION

The consistent and accurate interpretation of mammograms is one of the most challenging tasks in radiology and even among expert breast imagers there are variations in interpretation[1,2]. Despite these difficulties facing practicing radiologists, all radiology residency graduates are expected to be competent in the interpretation of mammograms upon graduation[3]. However, performance data of new radiology residency graduates demonstrates that those without subspecialty training in breast imaging require several years of on the job experience before their practice patterns fall in line with approved expectations[4]. Survey data has also shown that radiology residents are not confident in their interpretative abilities and feel that only those radiologists with specialized breast imaging training should routinely interpret mammograms[5,6]. As a result, there is a need for improved radiology residency training in mammography in order for graduates to be competent in the interpretation of mammograms without requiring extensive "on the job training."

During the four years of radiology residency, residents are required to spend at least 12 weeks of dedicated time on breast imaging[7]. The responsibilities on these weeks are distributed between interpreting mammograms, ultrasound, tomosynthesis, and magnetic resonance imaging (MRI) studies, in addition to performing procedures and counseling patients. Radiology residency education in mammography has relied predominately on supervised interpretations of mammograms with expert breast imagers, as well as didactic lectures and independent study. While supervised learning has the potential to address individual resident strengths and weaknesses, the time demands of a busy academic practice and the large number of supervising radiologists may place limits on the comprehensiveness of the educational experience. Meanwhile, didactic lectures provide a one size fits all approach to education which will not be tailored to individual resident's needs. As a result, there is a strong demand for a systematic and

comprehensive educational toolset that can identify individual resident weaknesses in a timely fashion, so as to provide opportunities for improvement.

There are broad trends in error making patterns that apply to inexperienced radiology residents during the interpretation of mammograms, including higher rates of false positive interpretations[8] and slower interpretation times[9]. Furthermore, recent work has suggested that individual residents also have unique error making patterns[10-13]. Error-making patterns may be captured using statistical and machine learning models[10,11] which could then be applied to the development of personalized educational tools to target individual radiologist weaknesses[10,12]. These educational tools have the potential to provide more consistent educational goals across radiology residency programs, and ensure that all residents meet appropriate standards by applying quantitative measures for the trainees' performance and progress. However, to build successful error-making models that relate image properties to error, a comprehensive understanding of the image features that are predictive of error is needed. The presented study is a step in this direction.

The Breast Imaging-Reporting and Data System (BI-RADS), a product of the American College of Radiology, provides a standardized language that is universally accepted in the United States and many other countries for the interpretations of mammograms[14]. Through the use of BI-RADS terminology, radiologists and referring clinicians are able to easily communicate significant findings. In addition to a final assessment score which assesses the overall likelihood of malignancy, the BI-RADS atlas provides descriptors of breast, mass, and calcification features. These features have been shown to be independently predictive of malignancy[15-18]; however, the relationship between these BI-RADS features and trainee interpretive abilities is not well established. A previous pilot study by our group demonstrated that individual user

models can be generated to predict error in distinguishing malignant and benign masses (i.e. characterization) for radiologists in training for two BI-RADS features: breast density and mass margin[10]. The effect of BI-RADS breast and mass features on radiology trainee performance in the detection of masses in mammograms is unknown. Since the detection of an abnormality is the first step towards determining whether a patient needs additional workup, the ability to detect mammographic masses is of paramount importance to radiology residents.

The purpose of this study is to apply statistical modeling to identify error making patterns of radiology trainees in the detection of mammographic masses using the framework of BI-RADS. If patterns exist, they may help provide important insights into radiology resident education and aid in the development of educational materials. Furthermore, the constructed models can be used directly in adaptive educational systems to predict which cases will pose high difficulty to the trainees.

MATERIALS AND METHODS

Reader Study

In order to quantitatively assess the relationship between trainee performance in the detection of mammographic masses and BI-RADS descriptors, we conducted a reader study. An Institutional Review Board protocol was obtained for the study. Ten radiology trainees and 3 expert breast imaging radiologists from the Duke University Department of Radiology and the University of North Carolina at Chapel Hill Department of Radiology participated in the study. The expert interpreters were all fellowship trained in breast imaging with 7 to 14 years of experience. The trainees included three breast imaging novices with no formal training, and seven radiology residents with at least 4 weeks of formal breast imaging training.

Each participant reviewed 100 mammographic cases in this reader study. Each case consisted of bilateral craniocaudal (CC) and medial lateral oblique (MLO) views. The data set consisted of normal, benign, and malignant cases. The abnormal (benign and malignant) cases contained biopsy-proven masses. Since the focus of the study was on mammographic masses, the readers were instructed to ignore calcifications. Each participant reviewed the images on a dedicated research workstation. For each case, the participant clicked on any actionable abnormalities identified on each view. An abnormality was considered actionable if the reader felt that additional workup was needed, which would include additional imaging or biopsy.

Additionally, for each actionable case, three experts were asked to record breast and mass features according to the BI-RADS lexicon. The breast features were breast density (fatty, scattered fibroglandular, heterogeneously dense, or extremely dense) and the presence of axillary lymph nodes (none, normal, or abnormal). The mass features were mass shape (round, oval, lobular, or irregular), mass margin (circumscribed, microlobulated, obscured, indistinct, or spiculated), and mass density (high, equal, low, or fat containing) as well as mass location including depth (anterior, middle, or posterior) and clock face (1 through 12).

Data Analysis

The expert interpretations were used to establish the ground truth. If at least two out of three experts assessed a case as actionable, then the final assessment was considered positive; otherwise, it was considered negative. Further analysis was performed on the 78 cases that the experts deemed positive. We assumed that a trainee made an error if his or her assessment was negative (i.e. actionable) and the ground truth was positive. Therefore, we focused only on the errors in the detection of masses (i.e. false negative errors). While many mammography studies consider error as a difference between the reader's assessment and the biopsy results, our goal

was to test how closely the resident's detection skills compare to expert reader's interpretations. Thus, we decided to use the expert interpretations as the ground truth for our study.

For each positive case, a final feature value was assigned for each breast and mass feature based on the assessments of the experts. If two or three of the experts agreed on a feature then that value was used. If all three experts disagreed, then the median value was used. In cases when not all experts reported values, if only two values were assigned then a randomly selected value from the two values was used. If only one value was assigned, then that value was used.

A statistical model was constructed for each trainee to capture error making patterns for the trainees. Specifically, the model (a classifier) used the breast mass features as independent variables (input) and occurrence of error as output. To construct the model, we used multinomial logistic regression and specifically the *mnrfit* function in the MATLAB software. The model was developed individually for each trainee using only that trainee's error-making data

To evaluate the predictive performance of the model, we first applied the leave-one-out cross validation scheme to calculate the prediction of likelihood of error by the model for each trainee and each case. Then, we used area under the receiver operating characteristic (ROC) curve. We calculated the area under the ROC curve (AUC) using the *prediction* and *performance* functions in the *ROCR* package in R statistical software (R Development Core Team). The AUCs were calculated separately for each trainee and then averaged across the trainees. Furthermore we applied the Jackknife procedure to estimate the confidence intervals for the AUC as well as test whether the mean AUC was significantly higher than 0.5 [19]. In the ROC analysis, the AUC value of 0.5 represents a "chance" (or "flip of a coin") prediction and value higher than 0.5 represents predictions better than chance. In the Jackknife procedure, one case was removed at a time and

the remaining cases were used to calculate the average AUC. We used in-house software written in R to conduct the Jackknife procedure.

RESULTS

The distribution of breast and mass features is shown in Table 1. This table also shows frequency of error (pooled across all residents) for specific values of all features. The majority of cases had either scattered fibroglandular (33/78, 42.3%) or heterogeneously dense (34/78, 43.6%) breast tissue. The majority of cases had normal axillary lymph nodes (61/78, 78.2%). The cases most commonly had masses with irregular shapes (34/78, 43.6%), indistinct margins (32/78, 41%), and equal density (68/78, 87.1%). The masses were most commonly located in the middle depth (38/78, 48.8%). The mode for the mass locations was 3 o'clock. The average detection error rate for all trainees was 31.8% (range: 16.7-43.6%, standard deviation: 10.1%).

The average performance of the models for the prediction of error measured by the AUC was 0.611 (95% Confidence Interval: 0.543 – 0.680, $p < 0.002$). The range of AUCs for the individual models was 0.502 to 0.739. This demonstrates that while the predictive accuracy of error-making models is far from perfect (as expected), there is a pattern in error making by the trainees (specifically in missing mammographic masses) in the form of a relationship between the BI-RADS features and error.

We also explored the individual patterns of error making for two trainees (trainees 2 and 8) who had similar detection accuracy (40% and 41%) and for whom the error-making model had good performance (AUCs of 0.739 and 0.720). The relationship between the BI-RADs feature values and frequency of error are shown in Figure 1. These graphs demonstrate unique trends for each trainee. The effect of mass shape on trainee detection accuracy demonstrates an inverse bell

curve for trainee 2, but a steady increase in detection accuracy for trainee 8. Meanwhile, mass depth appears to have little influence on Trainee 2 with a flat graph, but the detection performance for Trainee 8 shows an inverse bell curve. Finally, mass margin appears to have a shelf effect for Trainee 2, while demonstrating a bell curve shape for Trainee 8. These divergent patterns of detection accuracy demonstrate the importance of individualized modeling, as a one size fits all approach would likely not be able to capture the differences between trainees.

DISCUSSION

Our study demonstrates that a model of radiology trainee error making patterns in mass detection can be constructed using BI-RADS features as predictors. Radiology trainees do not make random errors, but rather commit errors that are related to distinct BI-RADS features. The models constructed and evaluated in this study not only show the existence of specific error-making patterns but also can be used in adaptive educational systems for the prediction of cases with high difficulty for each individual resident. This is also an important finding for the BI-RADS lexicon, which has a well-established association with a risk of malignancy, but has heretofore not been assessed in relation to error making patterns in mass detection for radiology trainees. By developing a predictive model using terminology that is well understood by clinical radiologists this allows for potentially better incorporation of subsequent educational tools into the mainstream educational toolset.

The next step in the process of incorporating these findings into practice would be the development of a personalized education infrastructure. We envision an environment where radiology trainees undergo testing at periodic intervals during their residency to assess for strengths and weaknesses. Once individual weaknesses have been identified, cases can be selected from a large categorized pool that have demonstrated a higher likelihood of error for the

individual trainee. The trainee can then review these cases in a simulated environment with feedback designed to correct error making habits. This can serve as a valuable supplement to the non-targeted educational methods that are currently being employed. It will also give residents the opportunity to recognize their deficiencies which can help direct their independent studies more efficiently. Additionally, by identifying which types of cases that residents struggle with interpreting, trends across residency programs can be identified which may result in curriculum changes in an effort to correct residency-wide deficiencies.

A short discussion regarding the significance of AUC values in this study is appropriate. While in many studies, an AUC of greater than 0.800 signifies a high significance, these expectations cannot be applied to a study like this one. The real world nature and heterogeneity of responses dulls the influence of computer modeling. There are multiple sources of trainee detection error that derive from random sources that cannot be adequately modeled. A trainee distracted by a noise in the next room or suffering from fatigue at the end of a long work day may skip a step in his or her search pattern and forget to look at a portion of the breast. While attempts are made during the study design to provide as real world an environment as possible, it is never truly possible to compensate for all potential random sources of error. As a result, the average AUC of 0.611 in this study, while not very high, does represent a statistically significant finding. Furthermore, while a model with AUC=0.611 would not be clinically useful for other tasks (such as computer-aided diagnosis), any model with AUC significantly higher than 0.5 is applicable in the context of computer-aided education since, when applied for selection of cases seen by the trainee, will increase the proportion of the training cases that a resident would find difficult and therefore it could improve the outcomes of their training. The quantity of material radiology residents are expected to learn is enormous, and the amount of time available is finite, and so any advantage that can be provided to residents will surely be warmly received.

This study has some limitations. First, our analysis was case-based rather than lesion-based i.e. allowing for example the situation in which the trainees identified a different lesion than experts (a known limitation of case-based analysis). However, this limitation does not question the validity of our results since our method was still able to predict cases of high and low difficulty for individual readers regardless of the potential noise introduced by the case-based approach. Future research could include more detailed, lesion-based analysis. Furthermore, the ground truth in this study was the expert majority opinion, since we wished to assess resident interpretation skills. However, it is possible that the expert majority opinion was incorrect and that instead the resident made the correct interpretation.

In summary, we were able to show that the BI-RADS lexicon is of use when building models of error making in the detection of mammographic masses. These models were shown useful for the identification of cases with a higher likelihood of error for individual trainees. This is a significant step towards building automatic adaptive educational systems for radiology.

CONCLUSION

One can construct a model that predicts whether a trainee will miss a mass in a mammogram better than chance using the BI-RADS lexicon features along with other standard clinical features as predictors. The proposed predictive models can be used to identify challenging cases for radiology residents and therefore potentially improve their educational outcomes.

ACKNOWLEDGEMENTS:

This research has been supported by the grant BC10444023 from Department of Defense Breast Cancer Research Program.

REFERENCES

1. T. Onega, M.L. Anderson, D.L. Miglioretti, et al., "Establishing a gold standard for test sets: variation in interpretive agreement of expert mammographers," Acad Radiol. 20(6), 731-739 (2013).
2. E. Lazarus, M.B. Mainiero, B. Schepps, S.L. Koelliker, L.S. Livingston, "BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value," Radiology. 239(2), 385-391 (2006).

3. D.L. Monticciolo, M. Rebner, C.M. Appleton, et al., "The ACR/Society of Breast Imaging Resident and Fellowship Training Curriculum for Breast Imaging, updated," J Am Coll Radiol. 10(3), 207-210 e204 (2013).

4. D.L. Miglioretti, C.C. Gard, P.A. Carney, et al., "When radiologists perform best: the learning curve in screening mammogram interpretation," Radiology. 253(3), 632-640 (2009).

5. L.W. Bassett, C. Bent, J.W. Sayre, R. Marzan, A. Verma, C. Porter, "Breast imaging training and attitudes: update survey of senior radiology residents," AJR. 197(1), 263-269 (2011).

6. L.W. Bassett, B.S. Monsees, R.A. Smith, et al., "Survey of radiology residents: breast imaging training and attitudes," Radiology. 227(3), 862-869 (2003).

7. E.A. Sickles, L.E. Philpotts, B.T. Parkinson, et al., "American College Of Radiology/Society of Breast Imaging curriculum for resident and fellow education in breast imaging," J Am Coll Radiol. 3(11), 879-884 (2006).

8. C.F. Nodine, H.L. Kundel, C. Mello-Thoms, et al., "How experience and training influence mammography expertise," Acad Radiol. 6(10), 575-585 (1999).

9. E.A. Krupinski, "Visual scanning patterns of radiologists searching mammograms," Acad Radiol. 3(2), 137-144 (1996).

10. M.A. Mazurowski, J.A. Baker, H.X. Barnhart, G.D. Tourassi, "Individualized computer-aided education in mammography based on user modeling: concept and preliminary experiments," Med Phys. 37(3), 1152-1160 (2010).

11. M.A. Mazurowski, H.X. Barnhart, J.A. Baker, G.D. Tourassi, "Identifying error-making patterns in assessment of mammographic BI-RADS descriptors among radiology residents using statistical pattern recognition," Acad Radiol. 19(7), 865-871 (2012).

12. S. Sun, P. Taylor, L. Wilkinson, L. Khoo. Individualised training to address variability of radiologists' performance. Proc SPIE 6917, Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment. San Diego, CA2008.

13. G. Tourassi, S. Voisin, V. Paquit, E. Krupinski, "Investigating the link between radiologists' gaze, diagnostic decision, and image content," J Am Med Inform Assoc. [Epub ahead of print] (2013).

14. C. D'Orsi, L. Bassett, W. Berg, et al.,"BI-RADS: Mammography," in *Breast Imaging Reporting and Data System: ACR BI-RADS - Breast Imaging Atlas, Vol. 1*, edited by C. D'Orsi, E. Mendelson, D. Ikeda, et al. (American College of Radiology, Reston, VA, 2003).

15. L. Liberman, A.F. Abramson, F.B. Squires, J.R. Glassman, E.A. Morris, D.D. Dershaw, "The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories," Am J Roentgenol. 171(1), 35-40 (1998).

16. J.M. Timmers, A.L. Verbeek, J. Inthout, R.M. Pijnappel, M.J. Broeders, G.J. den Heeten, "Breast cancer risk prediction model: a nomogram based on common mammographic screening findings," Eur Radiol. 23(9), 2413-2419 (2013).

17. Y. Wu, O. Alagoz, M.U. Ayvaci, et al., "A Comprehensive Methodology for Determining the Most Informative Mammographic Features," J Digit Imaging. [Epub ahead of print] (2013).

18. C.K. Bent, L.W. Bassett, C.J. D'Orsi, J.W. Sayre, "The positive predictive value of BI-RADS microcalcification descriptors and final assessment categories," Am J Roentgenol. 194(5), 1378-1383 (2010).

19. B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap,* 1st ed. (Chapman and Hall/CRC, Boca Raton, FL, 1994).

TABLES

Table 1: Distribution of breast and mass features and average error rate among all trainees.

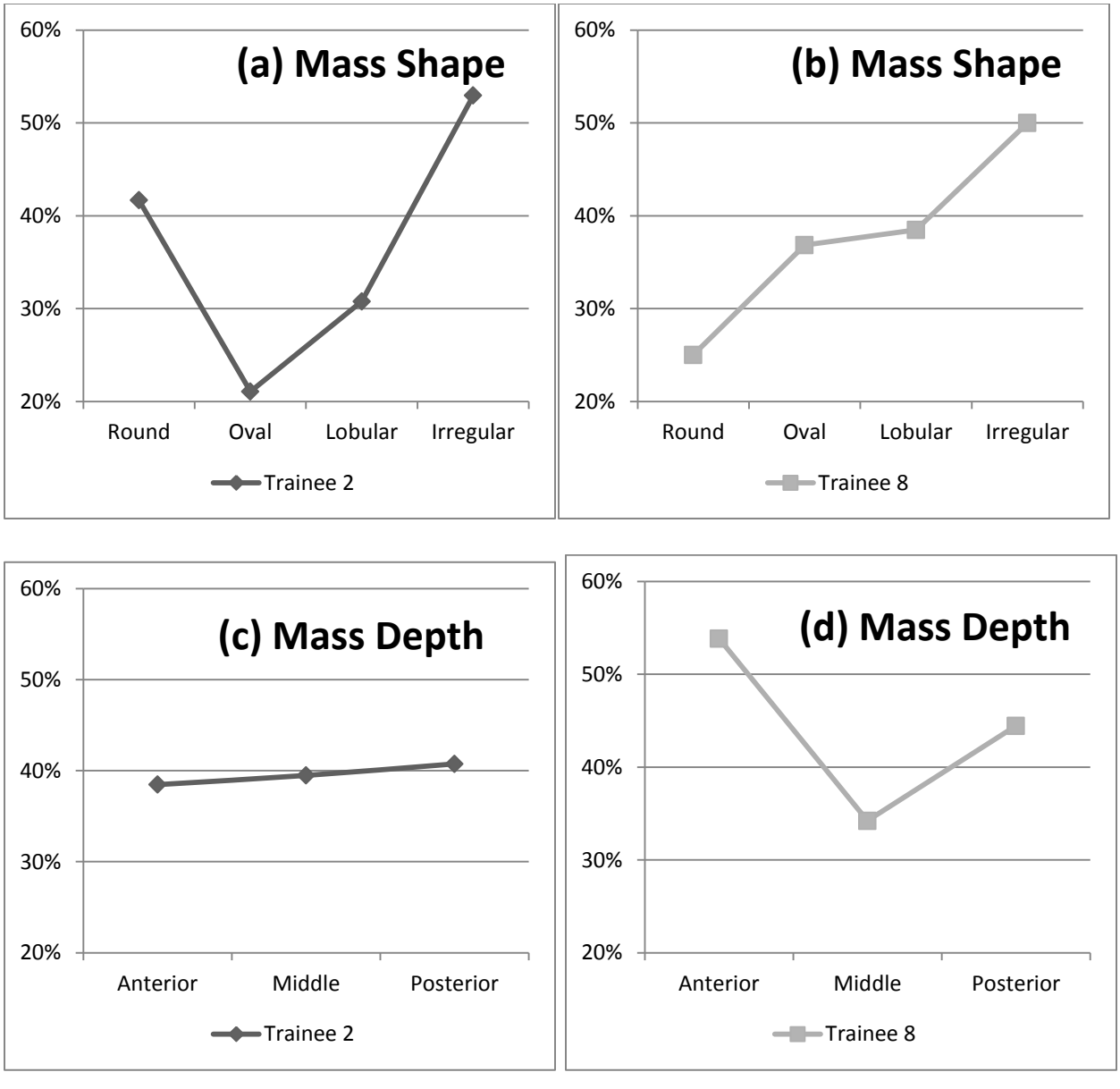
*Value for clock face denotes mode (average).

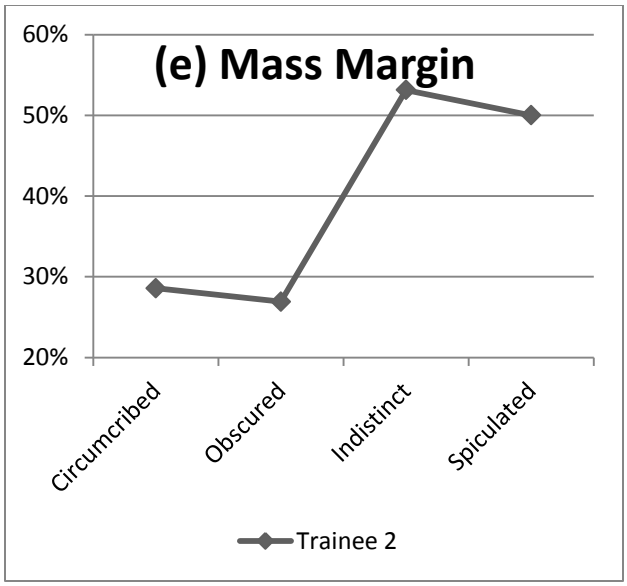
			Number (%)	Average error, %
Total			780 (100)	31.8
Breast features	Density	Fatty	6 (7.7)	16.7
		Scattered fibroglandular	33 (42.3)	27.9
		Heterogeneously dense	34 (43.6)	38.5
		Extremely dense	5 (6.4)	30.0
	Lymph nodes	None	15 (19.2)	28.7
		Normal	61 (78.2)	32.8
		Abnormal	2 (2.6)	25.0
Mass features	Shape	Round	12 (15.4)	22.5
		Oval	19 (24.4)	22.6
		Lobular	13 (16.7)	30.0
		Irregular	34 (43.6)	40.9
	Margin	Circumscribed	14 (17.9)	28.6
		Microlobulated	0 (0)	N/A

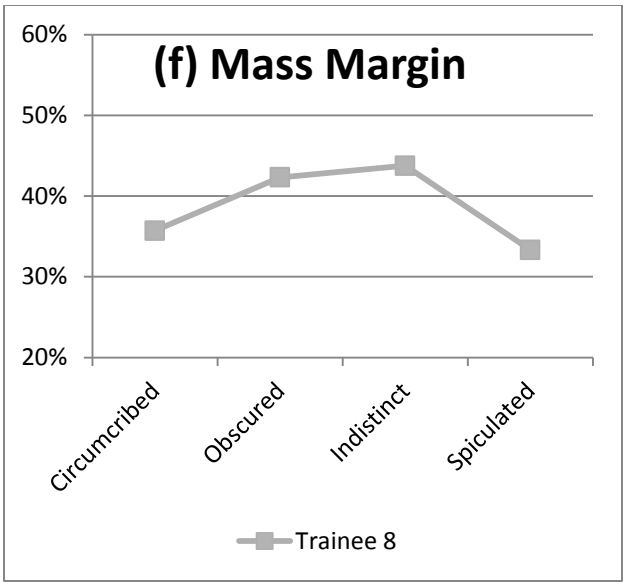
		Obscured	26 (33.3)	28.1
		Indistinct	32 (41.0)	36.3
		Spiculated	6 (7.7)	31.7
	Density	High	8 (10.3)	23.8
		Equal	68 (87.1)	32.8
		Low	2 (2.6)	30.0
		Fat containing	0 (0)	N/A
Mass location	Depth	Anterior	13 (16.7)	29.2
		Middle	38 (48.8)	27.9
		Posterior	27 (34.6)	38.5
Clock face			3 (6.8) *	N/A

FIGURES

Figure 1: Representative graphs of accuracy for trainees 2 and 8 for (a,b) mass shape, (c,d) mass depth, and (e,f) mass margin.. The average performance is 40% for trainee 2 and 41% for trainee 8. Of note, microlobulated margin was not included in the graph as there were no masses that met this criteria.




(e) Mass Margin
60%
50%
40%
30%
20%
Circumcribed
Obscured
Indistinct
Spiculated
Trainee 2


(f) Mass Margin
60%
50%
40%
30%
20%
Circumcribed
Obscured
Indistinct
Spiculated
Trainee 8

Difficulty of mammographic cases in the context of resident training: preliminary experimental data

Maciej A. Mazurowski[a,b]
[a] Carl E. Ravin Advanced Imaging Laboratories,
[b] Department of Radiology,
Duke University School of Medicine, Durham, NC

Abstract

We are currently developing an intelligent data-driven educational system for mammography. Since our system attempts to predict which cases will be difficult for the trainees, it is important to better understand the concept of case difficulty. While the concept of difficulty is central to our efforts on adaptive education, its importance extends to radiology education in general as well as to image perception research. In this study, we tested some hypotheses that related to difficulty. Specifically, we performed a preliminary reader study to evaluate relationship between the error rate (an objective measure of difficulty), individual assessment of case difficulty by a resident and expert's assessment of case difficulty (two subjective measures of difficulty). Furthermore, we investigated the relationship between individual and expert's assessment of difficulty and time that the residents took to interpret the case. Time taken to interpret a case by a resident related well with the individual assessment of difficulty but its relationship with the expert's assessment of difficulty was weaker. The analysis of the difficulty assessments showed that an increase in individual assessment of difficulty made by a resident relates well to an increase in his/her false positive errors but not to an increase in false negative errors. Interestingly, the expert's assessment of difficulty was related to false negative errors in the trainees but not to false positive errors. These results offer additional guidance in our efforts to construct an adaptive education system as well as provide insight into important aspects of radiology education in general.

Keywords: radiology education, reader study, mammography, difficulty of mammographic cases

1. INTRODUCTION

Mammography has been shown to play an important role in breast cancer detection. However, interpretation of mammograms is a very difficult process[1] and therefore efforts are needed to improve the interpretation skills of the radiologists working with mammograms. Education has been shown to play an important role in ensuring good performance of the radiologists interpreting mammograms[2,3].

We are currently developing a computerized training system for mammography that is using reading data previously collected for the trainees along with machine learning and statistical methods to improve the quality of the training[4-6]. Since the system attempts to identify cases that will be difficult for the trainees, the concept of case difficulty is very important.

In this study we test some basic hypotheses that pertain to this concept with focus on the relation between the difficulty assessed by a resident or an expert and actual errors. We believe that our conclusions, while laying a solid ground for our current work also inform the broader field of image perception and non-computer-based radiology education. Some previous studies analyzed performance of radiology trainees (such as studies by Nodine et al.[7,8]) but to our knowledge, none of them focused on the concept of difficulty.

2. METHODS

Four residents and one breast imaging expert participated in the study. An IRB protocol was obtained for the study. The participants came from two institutions: Duke University and University of North Carolina at Chapel Hill. Each participant read 100 mammographic cases. For each case, the participants identified the abnormality (if present), assessed likelihood of malignancy on a scale from 1 to 5, and assessed difficulty of the case on a scale from 1 to 5.

In this study, we focused on two questions related to case difficulty. The first question was: does individual (i.e. made by the resident interpreting the case) and expert's assessment of difficulty relate to the actual error rate (or equivalently to performance)? The actual error rate could be considered an ultimate but less accessible (than direct difficulty assessment) measure of difficulty. The second question was: does the assessment of difficulty for a case relate to time taken to interpret the case by the trainee? The answer to this question is intuitive for individual assessment (i.e. the residents take more time for cases that they perceive as difficult) but less intuitively simple for the expert's assessment.

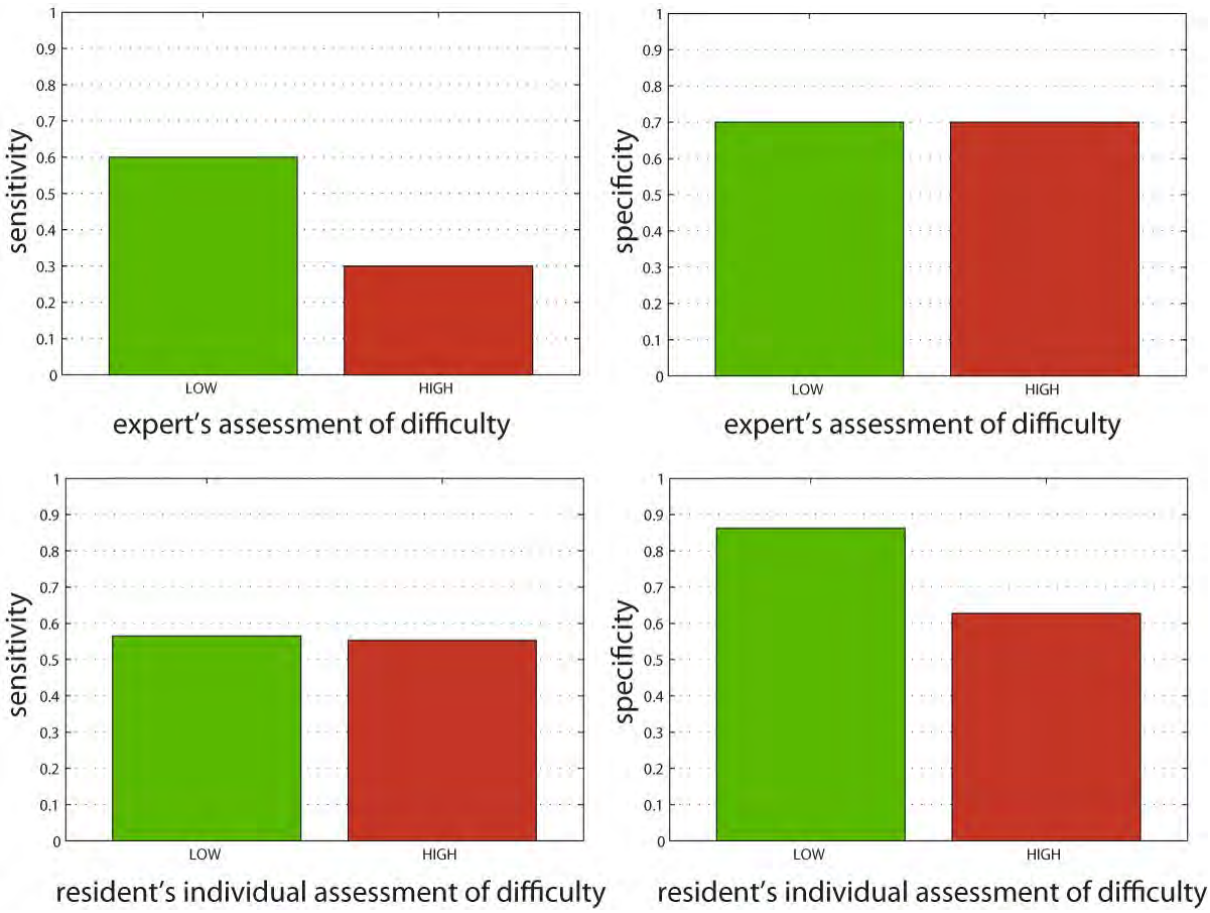


Figure 1 False positive and false negative rates. The top row shows errors stratified based on the individual assessment of difficulty by the residents. The bottom row shows the error rates stratified by the expert's assessment of difficulty.

To answer the first question, we stratified the cases based on individual and expert's assessment of difficulty and then examined the change in the quantities of interest. The quantities of interest were sensitivity and specificity which represent the rate of false negative and false positive errors respectively. We divided the cases into two groups: low difficulty and high difficulty. The threshold on difficulty assessment was established separately for the residents and for the expert in order to achieve the most balanced sets (the split was still highly imbalanced for the expert). To answer the second question (relationship of difficulty and interpretation time) we stratified the cases into three groups based on the individual resident's assessment and the expert's assessment. To established the three groups for both, the residents and the expert, the case was considered to be low difficulty if the difficulty assessment was 2 or less, medium difficulty if it was 3, and high difficulty if it was 4 or more. For each difficulty-based group in our analysis, pooled the results from all resident to see the general trends.

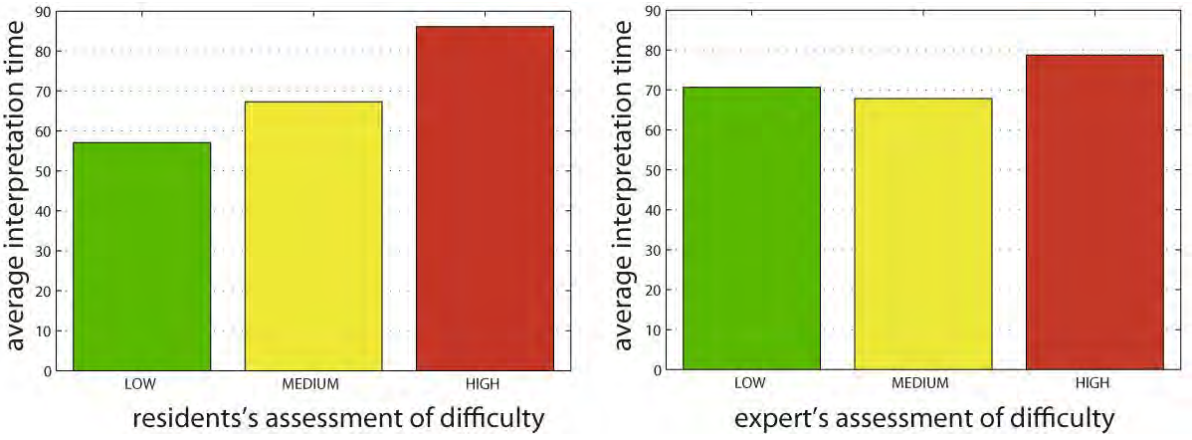


Figure 2 Average interpretation time by the residents for cases stratified by difficulty assessment. The plot on left shows average time for cases stratified by the individual difficulty assessment. The plot on the right shows average interpretation time for cases stratified by the expert's difficulty assessment.

The case was considered positive, if the assessment of likelihood of malignancy was 3 or more. The error was defined as a difference between the resident's decision (positive or negative) and the expert's decision. Such definition is useful in the context of education. The results are presented below.

3. RESULTS

Figure 1 presents the residents' sensitivity and specificity stratified by difficulty. In the upper row, the cases are stratified by the expert's assessment of difficulty and in the bottom row, the cases are stratified by the individual assessment of difficulty. It can be seen (bottom right plot) that the resident's assessment of difficulty is very well related to the actual false positive errors (lower specificity) made for the cases. This correlation, however, is not present for the false negative errors (upper right plot): the residents do not generally make more false negative errors for cases that they consider difficult. Regarding the difficulty assessment by the expert, it appears that it can be used to predict actual false negative errors (decrease in sensitivity with higher difficulty, see top left plot) but it is not very predictive of the false positive errors (no notable change in specificity, see bottom left plot).

Figure 2 shows the residents' interpretation time stratified to the assessment of difficulty. It can be seen that the residents generally take more time to interpret cases that they perceive difficult. However, interestingly the relationship between the assessment of difficulty by the expert and the time that the residents spent on the cases appears weaker.

4. CONCLUSIONS

In in this study, we tested some important hypotheses that relate to the concept of difficulty. Specifically, we evaluated relationship between the error rate (which could be considered an objective measure of difficulty), individual assessment of case difficulty by a resident and expert's assessment of case difficulty (i.e. subjective measures of difficulty). Furthermore, we investigated the relationship between individual and expert's assessment of difficulty and time that the residents took to interpret the case. We conducted a reader study where four residents and one expert read 100 mammographic cases each. The analysis of the results showed that the individual assessment of difficulty made by a resident relates well to the rate of false positives errors but not to the rate of false negative errors. Interestingly, the expert's assessment of difficulty predicted the rate of false negative errors but not the false positive errors. Time taken to interpret a case by a resident related well to the individual assessment of difficulty but the relationship with expert's assessment of difficulty was weaker. These results offer additional guidance in our efforts to construct an adaptive education system as well as provide insight into important aspects of radiology education in general.

5. ACKNOWLEDGEMENTS

This work was sponsored by the grant Grant BC10444023 from the Department of Defense Breast Cancer Research Program.

6. REFERENCES

1. Martin, J.E. & Moskowitz, M. Breast cancer missed by mammography. *American Journal of Roentgenology* **132**, 737-739 (1979).
2. Linver, M.*, et al.* Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. *Radiology* **184**, 39-43 (1992).
3. Leung, J.W.T.*, et al.* Performance parameters for screening and diagnostic mammography in a community practice: Are there differences between specialists and general radiologists? *American Journal of Roentgenology* **188**, 236-241 (2007).
4. Mazurowski, M.A., Baker, J.A., Barnhart, H.X. & Tourassi, G.D. Individualized computer-aided education in mammography based on user modeling: Concept and preliminary experiments. *Medical physics* **37**, 1152–1160 (2010).
5. Mazurowski, M.A. & Tourassi, G.D. Exploring the potential of collaborative filtering for user-adaptive mammography education. in *Biomedical Sciences and Engineering Conference (BSEC), 2011* 1-4 (IEEE, 2011).
6. Mazurowski, M.A., Barnhart, H.X., Baker, J.A. & Tourassi, G.D. Identifying Error-making Patterns in Assessment of Mammographic BI-RADS Descriptors among Radiology Residents Using Statistical Pattern Recognition. *Academic Radiology **19***, 865–871 (2012).

7. Nodine, C.F., Mello-Thoms, C., Kundel, H.L. & Weinstein, S.P. Time course of perception and decision making during mammographic interpretation. *American Journal of Roentgenology* **179**, 917-923 (2002).
8. Nodine, C.F.*, et al.* How experience and training influence mammography expertise. *Academic radiology* **6**, 575-585 (1999).